Discriminative Vision-Based Recovery and Recognition of Human Motion

Discriminative Vision-Based Recovery and Recognition of Human Motion

Ronald Poppe

9 789036 528108

CTIT Dissertation Series No. 09-136
Center for Telematics and Information Technology (CTIT)
P.O. Box 217, 7500 AE Enschede, The Netherlands

Ronald Poppe

Discriminative Vision-Based Recovery and Recognition of Human Motion

Ronald Poppe

PhD dissertation committee:

Chairman and Secretary:

Prof. dr. ir. Ton J. Mouthaan, University of Twente, NL

Promotor:

Prof. dr. ir. Anton Nijholt, University of Twente, NL

Assistant-promotor:

Dr. Mannes Poel, University of Twente, NL

Members:

Prof. dr. Hamid K. Aghajan, Stanford University, USA

Prof. dr. Dariu M. Gavrila, University of Amsterdam, NL and Daimler R&D, DE

Dr. Michael S. Lew, Leiden University, NL

Prof. dr. Maja Pantic, Imperial College, UK and University of Twente, NL

Prof. dr. Léon J.M. Rothkrantz, Delft University, NL and Defence Academy, NL

Dr. Raymond N.J. Veldhuis, University of Twente, NL

Human Media Interaction group

The research reported in this dissertation has been carried out at the Human Media Interaction group of the University of Twente.

CTIT Dissertation Series No. 09-136

Center for Telematics and Information Technology (CTIT)

P.O. Box 217, 7500 AE Enschede, NL

BSIK ICIS/CHIM

The research reported in this thesis has been carried out in the ICIS (Interactive Collaborative Information Systems) project. ICIS is sponsored by the Dutch government under contract BSIK03024.

SIKS Dissertation Series No. 2009-07

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-90-365-2810-8

ISSN: 1381-3617, number 09-136

Discriminative Vision-Based Recovery and Recognition of Human Motion

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee
to be publicly defended
on Thursday, April 2, 2009 at 16:45

by

Ronald Walter Poppe

born on May 21, 1980

in Tilburg, The Netherlands

This thesis has been approved by:
Prof. dr. ir. Anton Nijholt, University of Twente, NL (promotor)
Dr. Mannes Poel, University of Twente, NL (assistant-promotor)

ISBN: 978-90-365-2810-8

Acknowledgements

This dissertation has been approved by my promotor and co-promotor and, at last, I have come to appreciate it as well. When the task is to work on computer vision topics in a group that is highly multi-disciplinary, it's foolish to expect a paved road. Indeed, many of the insights and work reported in this dissertation have been obtained the hard way. While this may sound like regret, I've enjoyed the freedom of exploring and defining my research direction. Having enthusiastic people around working on different topics is stimulating and leads to new ideas that go beyond the scope of the main topic. It is those collaborations that have given an extra dimension to my four years as a PhD student and have shaped me more as a researcher. I won't deny that it took some time before I found a balance between the work that led to this dissertation and the 'distractions' but I guess that too was part of the learning process.

The domain of vision-based human motion analysis is very active as witnessed by the large amount of referenced literature. There have been rumors that I've used reading as an excuse to avoid doing 'the real work'. I admit that I got carried away with reading every once in a while but it paid off in insights, for me and hopefully for those who will read the literature overviews. The fast pace of the field has been frustrating at times when the very ideas I was working on were just about to be published (usually with better results!). On the other hand, it has been stimulating to work in a field that progresses so quickly. I can honestly say that I fully believe the claims that promise a whole range of interesting and life-improving applications. Sometimes, as I picture the people cheering and dancing as they use those applications that wouldn't have been possible without 'our' work, I realize it was worth the effort. (Hopefully, their happiness will be recorded for further analysis, to keep us busy.)

Luckily, this dissertation is a team effort rather than the result of one person's work, although I had to type the whole thing myself. First, I thank Mannes Poel for encouraging me to start a PhD, for being patient and for being critical about my work. I've very much appreciated his pleasant way of supervising, even though I've never mentioned this explicitly. Anton Nijholt has been an invaluable source of insights that any PhD student would love to know but which have never been written down. Moreover, I truly admire his dedication to manage a group as diverse as the Human Media Group. I'm looking forward to our continued collaboration. Also, I'd like to thank my dissertation committee for finding the time to read and comment on this work. Dariu Gavrila also provided useful feedback during earlier stages of my PhD period. Hamid Aghajan and Maja Pantic provided me with opportunities to present my work and get to know the community, for which I thank them.

The work in this dissertation has largely been shaped by colleagues 'in the field' and the discussions I had with them. I'm grateful to Leonid Sigal for making available the HumanEva dataset that has been used throughout the work, and for insightful discussions. Lena Gorelick provided the Weizmann human action dataset, including missing sequences, on short notice. Liefeng Bo, Daniel Weinland and Carl Henrik Ek provided interesting comments. Also, I thank the authors who allowed me to reuse their figures in this dissertation. Most of this work has been carried out within the context of the ICIS project. Even though our topics were far apart, I've valued the discussions with my colleagues in the CHIM cluster.

I've appreciated the lively and stimulating atmosphere at the Human Media Interaction group. The lunches, discussions at the coffee machine and evenings out: I've enjoyed all of them. Several persons deserve a special mentioning. I'd like to thank Dennis and Rutger for stimulating discussions, advice and many collaborations on projects, papers and student supervision. Moreover, they have been a stable source of entertainment over the years. Also, my roommates Yulia and Yujia are to be praised and thanked for bearing with me and for all nice conversations about random topics. Conversations and collaborations with Dirk were always interesting and entertaining. Also, I've enjoyed discussing work and life with Nataša, Trung and Khiet. Lynn deserves a medal for her stringent type-checking. Finally, Charlotte and Alice were always helpful in taking care of administrative matters.

My friends have been a great source of support and distraction. With the risk of appearing too lazy to write down all names: thank you all! Special thanks to Kresten, Lara, Renze, Sebas, Stef and Tony for the winter sport holidays, barbecues and the way-too-few evenings out. Wouter did a great job in sharing experiences of PhD life. Conversations with Marcoen have always inspired me, even though we don't see each other that often. Edgar always encouraged me to do things 'in the mix'.

My (extended) family has kept me motivated throughout all those years. My parents deserve a special mention as they have taught me that it's no disgrace to work hard, which has proven to be a useful lesson. Their interest and support has been tremendously valuable. Last, but certainly not least, I'd like to thank my girlfriend Saskia for her patience, positive spirit and love. Even though becoming a PhD student was my own choice, she has always supported me, for which she deserves so much more than just a 'thank you'. Not only now, but also in the years to come.

Ronald Poppe
Enschede, March 2009

Contents

1

Introduction

1.1 Human motion analysis

The systematic analysis of human motion dates back at least to Aristotle. However, it was only in the late 19^{th} century that sequences of photographs could be recorded at sufficient speed for vision-based motion analysis. Pioneers in this field of chronophotography were Marey [209] and Muybridge [231]. Their recordings allowed for qualitative and quantitative analysis of human motion. The reader is referred to Klette and Tee [176] for a more detailed historic overview of human motion analysis.

The shift to automatic human motion analysis largely found its origin in the work by Johansson [156], who placed reflective markers on human joints. He showed that such a representation enabled human observers to recognize human action, gender and viewpoint. These compact representations of human motion also proved to be suitable for automatic recovery and recognition of human motion. However, since markers are usually absent in the image sequences, we focus on markerless, vision-based analysis of human movement.

The visual analysis of human motion comprises many aspects. In this thesis, we limit our focus to *human pose recovery* and *human action recognition*. The former is a regression task where the aim is to determine the locations or angles of key joints in the human body given an image of a human figure. The latter is the process of labelling image sequences with action labels, which is a classification task. Importantly, we do not consider the *interpretation* of the motion, which requires reasoning and is usually dependent on the specific application or application domain. For both the pose recovery and the action recognition task, we assume that the human figure in the image has been localized in a previous step. This process of *human detection* or *human localization* falls outside our scope. However, we briefly discuss this topic in Section 3.1.

The research context and focus of our work is further explained in Section 1.2. We will discuss the discriminative aspect of our work in Section 1.3. In Section 1.4, we summarize the contributions of the work described in this thesis. Finally, we present the outline of this thesis in Section 1.5.

1.2 Research context

The research in this thesis was carried out within the ICIS project (Interactive Collaborative Information Systems). The focus of this project is on the design, development and evaluation of computer-assisted crisis management systems. Situational awareness and automatic decision making are key topics within the project and humans play an important role in both these processes. The CHIM (Computational Human Interaction Modelling) cluster looks at humans and their interaction with a crisis management system. This interaction can be conscious, when humans actively control the interaction, for example, using gesture-based interfaces. Alternatively, the interaction can be unconscious. In this case, humans can be observed and the actions they perform can be used to increase the system's awareness of the situation. An example is the recognition of running persons from surveillance cameras.

Despite differences between conscious and unconscious human motion, both cases require *reliable* and *real-time* recovery of human poses and recognition of human actions. In this thesis, the focus is therefore on these criteria.

The application of our research is not limited to the crisis domain. Visual surveillance could also benefit from the work described in this thesis. This will enable recognition of malicious actions in shopping malls or parking lots or help in monitoring elderly people to enable them to live independently for a longer period of time. The work could also be used for human-computer interaction applications that require real-time visual processing. While our work is motivated by the crisis management domain, we do not restrict ourselves to a single domain. Rather, we use publicly available datasets. The use of these datasets allows for comparison with other work. Also, given the public nature of these datasets, the precise merits and limitations of our contributions may be more easily understood.

1.3 Discriminative pose recovery and action recognition

Human pose recovery and action recognition approaches can be either generative or discriminative. Generative approaches model the mapping from pose or action to image, usually by employing a human body model. This allows for generation of the observation, given a pose description or action class. Their advantage is that many parameters, such as body dimensions, visual appearance and viewpoint can be included in the model, which allows for more faithful reconstruction of the image or image representation. One drawback of generative approaches is the need for a reasonably accurate initial estimate. More importantly, generative approaches usually require many iterations to converge to the approximately correct solution. Given that model projection and projection-to-image matching are computationally demanding, generative approaches cannot operate in real-time without oversimplifying assumptions. This makes them unsuitable for the applications we focus on.

In contrast, discriminative approaches do not model the mapping from pose or action to image but rather learn the inverse of this mapping. The term discriminative is motivated by the fact that these approaches do not model a class of poses or actions, but instead learn how to distinguish between different poses or actions, conditioned

on the observation. This allows for direct evaluation of a mapping function.

For discriminative approaches, the mapping from observation to pose or action is complex and is usually learned from data. In the case of human pose recovery and action recognition, this requires observation-pose or observation-action pairs (in Chapter 6 we will also look at pose-action pairs). Consequently, we can only approximately recover and recognize those poses and actions that we use to learn the mapping. This makes discriminative approaches suitable only for domains that are constrained with respect to the poses, viewpoints and other variations that we explicitly want to deal with, and train on. For pose recovery, this implies that the number of parameters that we can recover is limited.

The great advantage of discriminative approaches is that, after learning the mapping offline, online inference can be performed with low computational cost. In fact, many discriminative works operate in real-time. This is of key importance for the interactive applications that we consider in our work. Therefore, in this thesis, we follow a discriminative approach for both pose recovery and action recognition.

A more thorough discussion of generative and discriminative approaches for pose recovery and action recognition, respectively, is presented in Chapters 2 and 4.

1.4 Contributions of this thesis

We focus on fast human pose recovery and human action recognition from images and video. As discussed previously, we take a discriminative approach in both tasks. We assume that the human figure in the image has been detected in a previous step. The extracted figure is further described in a more compact form: the *image representation*. For both the pose recovery and the action recognition task, we use an adaptation of histograms of oriented gradients (HOG). This grid-based image representation is compact but sufficiently informative, and is invariant to changes in translation, scale and lighting.

In this thesis, we make several contributions, which are summarized below. We consider the evaluation of our contributions as an important aspect. Therefore, we performed extensive experiments on publicly available datasets.

- We give an extensive overview of the state of the art in human pose recovery and human action recognition. We describe directions within each field and the advantages and limitations of different approaches, while focussing on recent work. (Chapters 2 and 4)

- We present an example-based approach to human pose recovery. In such an approach, the training examples are retained, and pose recovery of an unseen image is obtained by weighted interpolation of the poses associated with the closest visual examples. The performance of the approach does not rely on precise parameter setting and therefore allows for thorough investigation of the performance of the HOG descriptors. (Chapter 3)

- In realistic situations, partial occlusion of the human figure in the image is common. However, the recovery of human poses from partially occluded images

has been largely ignored. We adapt our example-based pose recovery approach to cope with partial observations, when these are predicted. We use the grid-based nature of the HOG descriptor to efficiently recover the pose using part of the image descriptor. Regardless of the area and type of occlusion, our adapted approach has the same computational complexity as the original example-based approach. (Section 3.3)

- To recognize human actions from image sequences, we describe each frame with a HOG descriptor. For each pair of action classes (e.g. walking or waving), we apply a common spatial pattern (CSP) transform on sequences of these descriptors. The transform uses differences in variance between the two classes to maximize separability. Each of the pair-wise discriminative functions softly votes into the two classes. After evaluation of all pair-wise functions, the class with the maximum voting mass is selected. Due to the simplicity of the functions, evaluation can be performed efficiently. (Chapter 5)

- We combine the example-based pose recovery approach with the CSP classifier to recognize human actions from sequences of recovered poses. Thanks to rotation normalization of the poses, we can train the action models independently of the viewpoint. Moreover, we can recognize actions from partially occluded image observations since we can deal with these occlusions in the pose recovery step. (Chapter 6)

1.5 Thesis outline

Human pose recovery and human action recognition are discussed in Part I and II of this thesis, respectively. Each part starts with an overview of the domain (Chapters 2 and 4), followed by our practical contributions to the fields (Chapters 3 and 5). In Part II, we also discuss how human actions can be recognized from recovered poses, thereby linking the two topics (Chapter 6).

In Part III, we summarize our main contributions and discuss the strengths and limitations of our approaches. Finally, we present avenues for future work (Chapter 7.3).

Part I

Human pose recovery

2

Human pose recovery: an overview

2.1 Introduction

Human body pose recovery, or pose estimation, is the process of estimating the configuration of body parts from sensor input. When poses are estimated over time, the term human motion analysis is used. Traditionally, motion capture systems require that markers are attached to the body. These systems have some major drawbacks as they are obtrusive, expensive and impractical in applications in which the observed humans are not necessarily cooperative. As such, many applications, especially in surveillance and human-computer interaction (HCI), would benefit from a solution that is markerless. Vision-based motion capture systems attempt to provide such a solution using cameras as sensors. Over the last two decades, this topic has received much interest and it continues to be an active research domain. In this overview, we summarize the characteristics of and challenges presented by markerless vision-based human motion analysis. We discuss recent literature but we do not intend to give complete coverage to all work.

2.1.1 Scope of this overview

Human motion analysis is a broad concept. In theory, as many details as the human body can exhibit could be estimated, such as facial movement and movement of the fingers. In this overview, we focus on large body parts (torso, head, limbs). We limit ourselves to estimating body part configurations over time and not recognition of the movement. Action recognition, which is interpreting the movement over time, is not discussed in this overview. See Chapter 4 for an overview of action recognition literature. Surveys on gesture recognition appear in [83; 266]. For some applications, the positioning of individual body parts is not important. Instead, the entire body is tracked as a single object, and such applications are termed human tracking or detection. This is often a preprocessing step for human motion analysis, and we will not discuss the topic in this overview. A brief discussion appears in Section 3.1. In the remainder of this section, we summarize past surveys and taxonomies, and describe the taxonomy that is used throughout this overview.

2.1.2 Surveys and taxonomies

Within the domain of human motion analysis, several surveys have been written, each with a specific focus and taxonomy. Gavrila [108] divides research into 2D and 3D approaches. 2D approaches are further subdivided into approaches with or without the explicit use of shape models. Aggarwal and Cai [5] use a taxonomy with three categories: body structure analysis, tracking and recognition. Body structure analysis is essentially pose estimation and is split up into model-based and model-free, depending upon whether *a priori* information about the object shape is employed. A taxonomy for tracking is divided into single and multiple perspectives. Moeslund *et al*. [222] use a taxonomy based on subsequent phases in the pose estimation process: initialization, tracking, pose estimation and recognition. Wang *et al*. [373] use a taxonomy similar to [5]: human detection, human tracking and human behavior understanding. Tracking is subdivided into model-based, region-based, active contour-based and feature-based. Wang and Singh [372] identify two phases in the process of computational analysis of human movement: tracking and motion analysis. Tracking is discussed for hands, head and full bodies. Forsyth *et al*. [97] discuss tracking and animation approaches dealing with human motion.

Currently, we see some new directions of research such as combining top-down and bottom-up models, particle filtering algorithms for tracking, and model-free approaches. We feel that many of these trends cannot be discussed appropriately within the taxonomies mentioned above. We observe that studies can be divided into two main classes: model-based and model-free approaches. Model-based approaches employ an *a priori* human body. The pose estimation process consists of modeling and estimation. Modeling is the construction of the likelihood function, taking into account the camera model, the image descriptors, human body model, matching function and (physical) constraints. We discuss the modeling process in detail in Section 2.2. Estimation is concerned with finding the most likely pose given the likelihood function. The estimation process is discussed in Section 2.3. Model-free approaches do not assume an *a priori* human body model but implicitly model variations in pose configuration, body shape, camera viewpoint and appearance. Due to their different nature in both modeling and estimation, we discuss them separately in Section 2.4. We conclude with a discussion of open challenges and promising directions of research. An earlier version of this overview appeared as [273].

Note that often the terms generative and discriminative are used. Discriminative approaches directly approximate the mapping from image to pose space, usually without using a human body model. Generative approaches are able to generate the input given a pose representation, for which typically a human body model is used. However, several works approximate this mapping functionally, thus without using a body model. Consequently, the modeling phase is significantly different. Therefore, we use the classes model-based and model-free instead.

2.2 Modeling

The goal of the modeling phase is to construct the function that gives the likelihood of an input image, given a set of parameters. These parameters include body config-

uration parameters, body shape and appearance parameters and camera viewpoint. Some of these parameters are assumed to be known in advance, for example a fixed camera viewpoint or known body part lengths. Estimating a smaller number of parameters makes the search for the optimal model instantiation more tractable but also poses limitations on the visual input that can be analyzed. Note that the relation between pose and observation is multivalued, in both directions. Due to the variations between people in shape and appearance, and a different camera viewpoint and environment, the same pose can have many different observations. Also, different poses can result in the same observation. Since the observation is a projection (or combination of projections when multiple cameras are deployed) of the real world, information is lost. When only a single camera is used, depth ambiguities can occur. Also, because the visual resolution of the observations is limited, small changes in pose can go unnoticed.

Model-based approaches use a human body model, which includes the kinematic structure and the body dimensions. In addition, a function is used that describes how the human body appears in the image domain, given the model's parameters. Human body models are described in Section 2.2.1.

Instead of using the original visual input, the image is often described in terms of edges, color regions or silhouettes. A matching function between visual input and the generated appearance of the human body model is needed to evaluate how well the model instantiation explains the visual input. Image descriptors and matching functions are described in Section 2.2.2. Other factors that influence the construction of the likelihood function are the camera parameters (Section 2.2.3) and environment settings (Section 2.2.4).

2.2.1 Human body models

Human body models describe the kinematic properties of the body (the skeleton) as well as the shape and appearance (the flesh and skin). We discuss these below.

2.2.1.1 Kinematic models

Most of the kinematic models describe the human body as a tree, consisting of segments that are linked by joints. Every joint contains a number of degrees of freedom (DOF), indicating in how many directions the joint can move. All DOF in the body model together form the pose representation. These models can be described in either 2D or 3D.

2D models are suitable for motion parallel to the image plane. Ju *et al*. [159] and Haritaoglu *et al*. [125] use a so-called Cardboard model in which the limbs are modeled as planar patches. Each segment has 7 parameters that allow it to rotate and scale according to the 3D motion. In [140], an extra patch width parameter was added to account for scaling during in-plane motion. In [2; 47], the human body is described by a 2D scaled prismatic model [227]. These models have fewer parameters and enforce 2D constraints on figure motion that are consistent with an underlying 3D kinematic model. But despite their success in capturing fronto-parallel human movement, the inability to encode joint angle limits and self-intersection constraints

renders 2D models unsuitable for tracking more complex movement.

3D models allow a maximum of three (orthogonal) rotations per joint. For each of the rotations individually, kinematic constraints can be imposed. Instead of segments that are linked with zero-displacement, Kakadiaris and Metaxas [163] model the connection by constraints on the limb ends. In a similar fashion, Sigal *et al*. [325] model the relationships between body parts as conditional probability distributions. Bregler *et al*. [41] introduce a twist motion model and exponential maps which simplify the relation between image motion and model motion. The kinematic DOF can be recovered robustly by solving simple linear systems under scaled orthogonal projection.

The parameters of the kinematic model such as limb lengths are sometimes assumed fixed. However, due to the large variability among people, this will lead to inaccurate pose estimations. Alternatively, these parameters can be recovered in an initialization step where the observed person is to adopt a specified pose [21; 45]. While this approach works well for many applications, it restricts use in surveillance or automatic annotation systems. Online adjustment of these parameters is possible by relying on statistical priors [115] or specific but common key poses [24; 54].

The number of DOF that are recovered varies between studies. In some studies, a mere 10 DOF are recovered in the upper body. Other studies estimate full-body poses with no less than 50 DOF. But even for a model with a limited number of DOF and a coarse resolution in (discrete) parameter space, the number of possible poses is very high. Applying kinematic constraints is an effective way of pruning the pose space by eliminating infeasible poses. Typical constraints are joint angle limits [66; 369] and limits on angular velocity and acceleration [399].

2.2.1.2 Shape models

Apart from the kinematic structure, the human shape is also modeled. Segments in 2D models can be described as rectangular or trapezoid-shaped patches, such as the Cardboard model [159] (see Fig. 2.1(a)). Segments in 3D models are either volumetric or surface-based. Volumetric shapes depend on only a few parameters. Commonly used models are spheres [257], cylinders [129; 295; 318] or tapered super-quadrics [63; 110; 171] (see Fig. 2.1(b)). Instead of modeling each segment as a separate rigid shape, surface-based models often employ a single surface for the entire human body [7; 13; 45] (see Fig. 2.1(c)). These models typically consist of a mesh of polygons that is deformed by changes to the underlying kinematic structure [20; 38; 162]. Plänkers and Fua [270] use a more complex body shape model, consisting of three layers: kinematic model, metaballs (soft objects) and a polygonal skin surface. When using 3D shape models, constraints can be introduced to prevent volume overlap of body parts [330].

Shape models can be assumed known or determined based on the observations. In several cases, the shape parameters are recovered jointly with the pose instantiation. Cheung *et al*. [51] and Mikić *et al*. [215] use a number of cameras and recover segment shape and joint positions by looking at motion of individual points. The parameters of a statistical model of human body shape [13] are estimated by Bălan *et al*. and Mündermann *et al*. [18; 230; 320]. Rosenhahn *et al*. [301] model additional clothing parameters for the lower body.

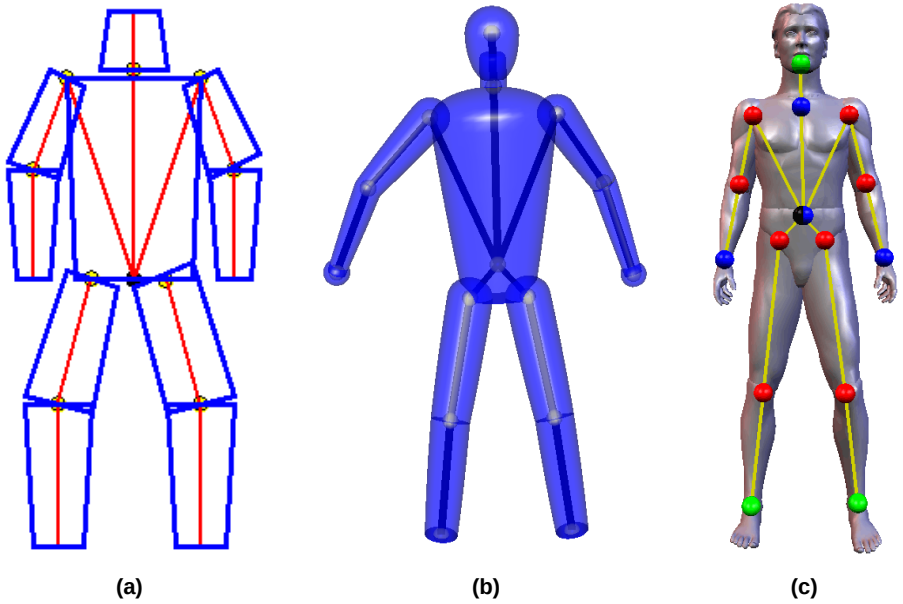


Figure 2.1: Human shape models with kinematic model. (**a**) 2D model (reprinted from [140], © IEEE 2002) (**b**) 3D volumetric model consisting of superquadrics (reprinted from [171], © Elsevier, 2006) (**c**) 3D surface model (reprinted from [45], © ACM, Inc., 2003)

To evaluate the likeliness of the model instantiation given the image, a function is required that describes how the instantiation appears in the image domain. An appropriate distance measure between synthesized model projection and image observation gives the likeliness of the model instantiation. We describe model appearance in the image domain and the matching functions in the next section.

2.2.2 Image descriptors

The appearance of people in images varies due to different clothing and lighting conditions. Since we focus on the recovery of the kinematic configuration of a person, we would like to generalize over these kinds of variation. Part of this generalization can be handled in the image domain by extracting invariant image descriptors rather than taking the original image. For synthesis, this means that we do not need complete knowledge about how a model instantiation appears in the image domain. Often used image descriptors include silhouettes, edges, 3D information, motion and color.

2.2.2.1 Silhouettes and contours

Silhouettes and contours (silhouette outlines) can be extracted relatively robustly from images when backgrounds are reasonably static. In older studies, backgrounds were often assumed to be different in appearance from the person. This eliminates the need to estimate environment parameters. Silhouettes are insensitive to variations in appearance such as color and texture, and encode a great deal of information to help recover 3D poses. However, performance is limited due to artifacts such as

shadows and noisy background segmentation, and it is often difficult or impossible to recover certain DOF due to the lack of depth information (see Fig. 2.2). Area overlap is commonly used as a distance measure between observed and synthesized silhouettes. In model-free approaches, silhouettes are encoded using central moments [40] or Hu moments [298]. Contours can be encoded using a combination of turning angle metric and Chamfer distance [132] or shape contexts [23], and can be compared based on deformation cost [225].

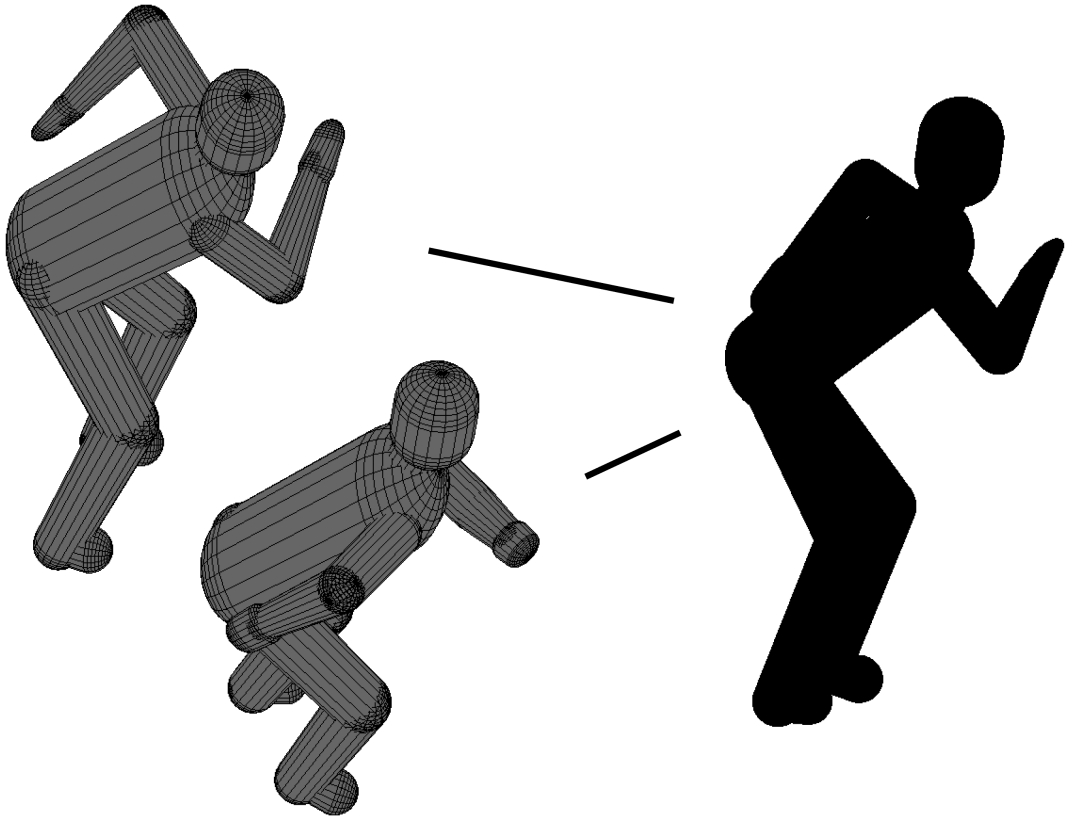


Figure 2.2: Depth ambiguities when using silhouettes from a single view [132] (© IEEE, 2004)

2.2.2.2 Edges

Edges appear in the image when there is a substantial difference in intensity at different sides of the image location. Edges can be extracted robustly and at low cost. They are, to some extent, invariant to lighting conditions, but are unsuitable when dealing with cluttered backgrounds or textured clothing. Therefore, edges are usually located within an extracted silhouette [163; 295; 369] or within a projection of a human model [72]. Matching functions take into account the normalized distance between a model's synthesized edges and the closest edge found in the image (Chamfer distance). Rohr [295] uses edge lines instead of edges to partially eliminate silhouette noise. A distance measure based on difference in line segment length, center position and angle is applied. When there is low contrast between background and foreground, or between human body parts, edges responses will generally be low.

Also, fast camera movement or relative movement of body parts can cause motion blur which hinders the robust extraction of edges.

2.2.2.3 3D information

Edges and silhouettes lack depth information, at least when only a single view is used. This also makes it hard to detect self-occlusions. When multiple views are present, each of them can be used individually to evaluate the model instantiation (e.g. [18; 63; 300]). Bălan *et al*. [16] additionally use shadows to match their projections to.

When two calibrated cameras are used, a depth map can be obtained by using stereometry. Corresponding points are sought in both views and the depths of these points are calculated using triangulation. This approach has been taken by Plänkers and Fua [270] and Haritaoglu *et al*. [125]. Stereo is also used by Jojic *et al*. [157], with the optional aid of projected light patterns. Matching functions are based on the volume overlap or average closest point distance.

When multiple cameras are used, a 3D reconstruction can be created from silhouettes that are extracted in each view individually. Two common techniques are volume intersection [38] or a voxel-based approach [51; 213; 215]. Such reconstructions can be matched against a model instantiation using volume overlap or measuring the average distance to the surface [130].

The accuracy of the 3D construction relies heavily on robust silhouette extraction. Moreover, the number of available views determines the level of detail. Several works have used additional image features to refine the 3D pose. Vlasic *et al*. [367] use details from individual silhouettes to align body parts, Aguiar *et al*. [6] and Starck and Hilton [335] use additional stereo and silhouette information, Aguiar *et al*. [7] use optical flow. While such refinements allow for accurate reconstruction of shape and pose, the computation time required prohibits their use in interactive applications.

For most of these works, tight-fitting clothes are assumed. Rosenhahn *et al*. [301] explicitly model clothing parameters for the lower-body. Bălan *et al*. [15] use a statistical body model and shape constraints over time. Also, skin color detection is used to find skin regions, that are assumed to fit the visual hull boundary. Ukita *et al*. [356] match a voxel model in a space that is constructed offline from examples.

2.2.2.4 Color and texture

Modeling the human body based on color or texture is inspired by the observation that the appearance of individual body parts remains substantially unchanged although the body may exhibit very different poses. The appearance of individual body parts can be described using Gaussian color distributions [392] or color histograms [283]. Roberts *et al*. [289] propose a 3D appearance model to overcome the problems with changing appearance due to clothing, illumination and rotations. They model body parts with truncated cylinders with surface patches described by a multi-modal color distribution. The appearance model is constructed on-line from monocular image streams. Skin color can be a good cue for finding head and hands. In [192], additional clothing parameters are used to model sleeve, hem and sock lengths.

2.2.2.5 Motion

Motion can be measured by taking the difference between two consecutive frames. The brightness of the pixels that are part of the person in the image are assumed to be constant. The pixel displacement in the image is termed optical flow and is used by Bregler *et al*. [41] and Ju *et al*. [159]. Sminchisescu and Triggs [330] use optical flow to construct an outlier map that is used to give weighting to the edges. Optical flow provides valuable information about the movement, which is independent of the appearance of the body. However, it proves more difficult to deal with cluttered, dynamic backgrounds and moving camera viewpoints.

2.2.2.6 Combination of descriptors

A likelihood function that takes into account a combination of descriptors proves to be more robust. Silhouette information can be combined with edges [66], optical flow [133] or color [51]. In [317], edges, ridges and motion are used. Filter responses for these image cues are learned from training data. Ramanan *et al*. [283] use edges and appearance cues. Care must be taken in constructing the likelihood function, especially when multiple image descriptors are used. Not unusually, a body part configuration that results in a low cost for one image descriptor will also result in a low cost for a second one. When the likelihood function simply multiplies the likelihood function for each image descriptor, this may lead to sharp peaks in the likelihood surface. This results in less efficient estimation.

2.2.3 Camera considerations

Monocular work [4; 318; 330] is appealing since for many applications only a single camera is available. When only a single view is used, self-occlusions and depth ambiguities can occur. Sminchisescu and Triggs [330] estimate that roughly one third of all DOF are almost unobservable. These DOF mainly correspond to motions perpendicular to the image plane, but also to rotations of near-cylindrical limbs about their axes. When multiple cameras are used, these DOF can still be observed. In general, there are two main approaches to use observations from multiple views. One is to search for features in each camera image separately and in a later stage combine the information to resolve ambiguities [63; 287; 300]. Wu and Aghajan take an opportunistic approach where rough estimates for body parts are determined from each view, and combined into a final estimate. Due to the limited amount of features that is collected, this approach drastically reduces bandwidth use. The second approach is to combine the information into a 3D reconstruction, as described before. When multiple cameras are used, calibration is an important requirement [350]. Instead of combining the views, Kakadiaris and Metaxas [163] use active viewpoint selection to determine which cameras are suitable for estimation.

While model-based approaches can, in theory, recover poses from any viewpoint, the vast majority of all works consider only the case where the camera is approximately at the height of the head. This is also true for model-free approaches, where the training set should account for the viewpoint.

Most studies assume a scaled orthographic projection which limits their use to distant observations, where perspective effects are small. Rogez *et al*. [292] remove the perspective effect in a preprocessing step.

2.2.4 Environment considerations

Most of the approaches described in this overview can handle only a single person at a time. Pose estimation of more than one person at the same time is difficult because of occlusions and possible interactions between the persons. However, Mittal *et al*. [221] were able to extract silhouettes of all persons in the scene using the M_2 tracker. A setup with five cameras provides the input for their method. The W^4S system [125] is able to track multiple persons and estimate their poses in outdoor scenes using stereo image pairs and appearance cues.

The results that are obtained are largely influenced by the complexity of the environment. Outdoor scenes are much more challenging due to the dynamic background and lighting conditions. Also, persons are often visible without occlusion by other objects. Only few works explicitly address partial occlusion, but rely on strong motion models [268; 276]. It remains a challenge to recover arbitrary poses of people under significant occlusion.

2.3 Estimation

The estimation process is concerned with finding the set of pose parameters that minimizes the error between observations and the projection of the human body model. This process proceeds in either top-down or in bottom-up fashion. Top-down approaches match a projection of the human body with the image observation directly, and iteratively refine the pose estimate. The top-down character is due to the hierarchical ordering of the human body model. Bottom-up approaches start by finding individual body parts, which are then assembled into a human body. Recent work combines these two classes. We discuss both classes and their combination in Section 2.3.1.

Recall that the goal of the modeling phase is to construct the pose likelihood function, given an image. The most likely pose estimate ideally corresponds to the global maximum of this function. However, the likelihood function often has many local maxima. Given the high dimensionality of the pose space, the search for the global maximum must be efficient. The speed of the pose recovery depends largely on the speed of the search strategy.

When images from multiple time instances are available, poses can be estimated over time as well. The pose estimates of the previous frame or frames can be used to give an initial estimate of the pose in the current frame. This process is called tracking or filtering, and is discussed in Section 2.3.2. Many methods are single-hypothesis approaches, which are deterministic in nature. Recent studies maintain multiple hypotheses, in either a deterministic or probabilistic fashion. This reduces the probability of getting stuck at a local maximum.

When predicting the current pose from past poses, a motion model is used. This can be a general (e.g. linear) model or be specific for a given action. It is often

observed that, for a given action, the movement of different body parts is strongly correlated and populates only a small volume in the space of possible body poses. Dimensionality reduction is often used to construct a lower-dimensional pose space that facilitates tracking at the cost of being restricted to a certain class of movements. We discuss these motion priors in Section 2.3.3.

Finally, we discuss the recovery of 3D poses from 2D pose representations in Section 2.3.4. While these approaches are strictly not within the scope of our survey, we discuss them briefly due to their close relation to many of the works described in this chapter.

2.3.1 Top-down and bottom-up estimation

There are two directions of model-based estimation: top-down and bottom-up. Recent work combines these approaches to benefit from the advantages of both, and is discussed in Section 2.3.1.3.

2.3.1.1 Top-down estimation

Top-down approaches match a projection of the human body with the image observation. Due to the flexibility of a projection function, top-down approaches can include many parameters and consequently explain the image observation accurately. This is termed an analysis-by-synthesis approach. The fitness of the match results in a likeliness score. A global search for the maximum score is often infeasible due to the high dimension of the pose space. Gall *et al*. [105] introduced a particle-based global optimization algorithm that shows resemblances to simulated annealing. While such an approach can be used to automatically initialize tracking, the efficiency in obtaining an accurate estimate is much lower compared to a local search around a close estimate. In Gall *et al*. [104], global optimization is combined with both tracking and local optimization, to yield a more efficient approach to obtain accurate pose estimation. In the case of local optimization, the *a posteriori* pose estimate is often found by applying gradient ascent on the likelihood surface [369]. Instead of performing this search in the pose (or parameter) space, Delamarre and Faugeras [63] iteratively minimize the discrepancy between between extracted silhouettes and the projected model. Local optimization is performed starting from a close estimate. This implies that (manual) initialization is needed. In a tracking approach, this is also true for the first frame.

To reduce the search in a high dimensional parameter space, Gavrila and Davis [110] use search-space decomposition. Poses are estimated in a hierarchical coarse-to-fine strategy, estimating the torso and head first and then working down the limbs. They further use a discrete pose representation, which results in a limited number of possible solutions per joint. Top-down estimation often causes problems with (self)occlusions, especially when search-space decomposition is used as errors can be propagated through the kinematic chain. An inaccurate estimation for the torso/head part can cause errors in estimating the orientation of body parts lower in the kinematic chain. To overcome this problem, Drummond and Cipolla [72] introduce constraints between linked body parts in the chain. This allows lower parts to effect parts higher

in the chain.

One important drawback of top-down approaches is the computational cost of forward rendering the human body model and calculating the distance between the rendered model and the image observation. Both of these processes are computationally expensive, and have to be performed at each iteration. When a sampling-based estimation approach is used for local optimization or tracking (see also Section 2.3.2), the number of samples is often too high to allow real-time pose recovery.

2.3.1.2 Bottom-up estimation

Bottom-up approaches are characterized by finding individual body parts and then assembling these into a human body. The assembling process takes into account physical constraints such as body part proximity. Bottom-up approaches have the advantage that no manual initialization is needed and can be used as an initialization for top-down approaches (see also Section 2.3.1.3). The body parts are usually described by 2D templates. Often, these templates produce many false positives, as there are often many limb-like regions in an image. Another drawback is the need for part detectors for most body parts, since missing information is likely to result in a less accurate pose estimate, unless pose priors are used. This requirement is difficult to meet as some limbs might have little image support when they are orthogonal to the image plane.

Felzenszwalb and Huttenlocher [89] model body parts as 2D appearance models. They use the concept of pictorial structures to model the coherence between body parts. An efficient dynamic programming algorithm is used to find an optimal solution in the tree of body configurations. Ronfard *et al*. [296] use the pictorial structures concept but replace the body part detectors by more complex ones that learn appearance models using Support Vector Machines. Ramanan *et al*. [283] automatically learn person-specific models of appearance, initially aided by parallel lines. Motion tracking is reduced to the problem of inference in a dynamic Bayesian network. The approach can (re)initialize automatically but tracking occasionally fails, especially for in-plane motion. Siddiqui and Medioni [316] use a tree model with encoded joint constraints. When traversing the tree in a bottom-up fashion, local optimal but sufficiently distinctive assemblies are maintained, thus drastically reducing the number of candidate poses.

In case of (self)occlusion, tree models generally have difficulty explaining the observation, which can lead to incorrect estimates. This issue can be overcome by learning multiple trees, for different combinations of limbs. Ioffe and Forsyth [146] use such a mixture of trees, where constraints between body parts are shared between different trees. Wang and Mori [382] use a similar idea, but apply boosting to discriminatively learn the tree models.

When using trees, only dependencies between body parts that are kinematically linked can be modeled. Trees are extended with correlations between body parts in [184] to enforce pose symmetry and balance. For walking, correlations between upper arm and leg swings are used, resulting in more robust pose estimations. A very similar approach has been taken by Ren *et al*. [288], who introduce arbitrary relations between body parts to model occlusion, scale, appearance and boundary smoothness. Pose estimation reduces to a shortest-path problem. Jiang *et al*. [155] explicitly dis-

tinguish between strong tree edges and weaker inter-part edges that model exclusion constraints. This allows them to infer the globally most likely pose more efficiently.

Sigal *et al*. [325] describe the human body as a graphical model where each node represents a parameterized body part (see Fig. 2.3(a)). Spatial constraints between body parts are modeled as arcs. Each node in the graph has an associated image likelihood function that models the probability of observing image measurements conditioned on the position and orientation of the part. Non-parametric belief propagation is used to infer the most likely pose. In [321; 124], temporal constraints are also taken into account, resulting in a tracking framework. Sigal and Black [323] use occlusion-sensitive image likelihoods which require relations between parts. This introduces loops in the graphical model, and approximate loopy belief propagation is used for inference. Gupta *et al*. [121] take a similar approach, but use observations from multiple views.

Instead of using limb detectors, Mori *et al*. [226] first perform image segmentation based on contour, shape and appearance cues. The segments are classified by body part locators for half-limbs and torso that are trained on image cues. From this partial configuration, missing body parts are found. The search space is pruned using global constraints, including body part proximity, relative widths and lengths and symmetry in color. Kuo *et al*. [182] cluster edge orientation, local motion and color to find clusters that could correspond to body parts. A 2D body model is then used to guide the clustering, leading to a iterative pose refinement. Ramanan [281] also refine the pose estimate iteratively, but does so by constructing more informative body part locators in each iteration. Ferrari *et al*. [93] extend this work to progressively reduce the search space by modeling background and employing temporal consistency. The above works can deal with a wide variety of human appearances, but generally produce less accurate pose estimates due to the lack of assumptions that is posed on the observation.

Demirdjian and Urtasun [64] discriminatively select a set of image patches. The pose density is approximated by kernels associated with the best-matching reference patches. Patches of approximately the size of a limb are used, which often match for a large range of poses. Poppe and Poel [276] therefore use templates of whole legs and arms, which implicitly encode 3D poses. They recover walking poses from various viewpoints under partial occlusions.

Instead of relying on appearance cues, Daubney *et al*. [61] use sparse motion features and determine the probability that a certain movement region belongs to a specific body part. They focus on walking motions, and determine the gait phase before refining the pose estimate.

2.3.1.3 Combined top-down and bottom-up estimation

By combining pure top-down and bottom-up approaches, the drawbacks of both can be alleviated. Automatic initialization can be achieved within a sound tracking framework.

Navaratnam *et al*. [236] use a search-space decomposition approach. Body parts lower in the kinematic chain are found using part detectors within an image region that is defined by their parent in the kinematic chain. This approach is computation-

ally less expensive but performance depends heavily on the individual part detectors.

Hua and Wu [138] incorporate bottom-up information in a graphical model of the human body, which encodes the observation likelihood of each body part, the spatial relations between them and the dynamics. A sampling-based approach is used to infer the most likely pose. Ramanan and Sminchisescu [284] use a conditional random field (CRF) to take into account a number of image features. They learn the parameters of the model from training data and, for a test image, maximize the likelihood for joint localization of all body parts. Kohli *et al*. [178] also use the CRF formulation, but introduce a pose-specific prior to aid the segmentation. Moreover, dynamic graph cuts are used for efficient inference.

Lee and Cohen [192] use part detectors and inverse kinematics to estimate part of the pose space. Bottom-up information is only used when available, eliminating the need for a part detector for each limb. The approach targets the drawbacks of a pure top-down approach, while still providing a flexible tracking framework. However, the bottom-up information in used in a fixed analytical way. This requires fixed segment lengths and prevents correct estimation of certain types of poses (e.g. poses where the elbow is higher than the hand). Proposal maps are introduced to facilitate the mapping from 2D observations to 3D pose space. Based on this work, Lee and Nevatia [193] focus on cluttered scenes and adopt a three-stage approach to subsequently find human bodies, their 2D body part locations and a 3D pose estimate.

2.3.2 Tracking

Estimating poses from frame to frame is termed tracking or filtering. It is used to ensure temporal coherence between poses over time and to provide an initial pose estimate. When it is assumed that the time between subsequent frames is small, the distance in body configuration is likely to be small as well. These configuration differences can be approximately linearly tracked, for example using a Kalman filter [162; 369]. Traditionally, tracking was aimed at maintaining a single hypothesis over time. However, ambiguity in the observation (e.g. when using silhouettes) causes the likelihood function to have multiple peaks. When only a single hypothesis is kept, there is the risk of selecting the wrong mode which causes the pose estimate to drift off. Recent work therefore propagates multiple hypotheses in time. Often, a sampling-based approach is taken. In some works, temporal coherence is achieved by minimizing pose changes over a sequence of frames in a batch approach. This section discusses multiple hypothesis tracking and batch approaches, respectively.

2.3.2.1 Multiple hypothesis tracking

To overcome the drift problem of single hypothesis tracking approaches, multiple hypotheses can be maintained. Cham and Rehg [47] use a set of Kalman filters to propagate multiple hypotheses. This results in more reliable motion tracking than with a single Kalman filter. Human motion is non-linear due to joint accelerations. However, Kalman filters are only suitable for tracking linear motion. Sampling-based approaches (particle filtering or Condensation [113; 147]) are able to track non-linear motion. In general, a number of particles is propagated in time using a model of

dynamics, including a noise component. Each particle has an associated weight that is updated according to the likelihood function. Configurations with a high likelihood are assigned a high weight. Since all weights sum up to one, the pose estimate is obtained by the weighted sum of all particles. (Or alternatively, the particle with the maximum weight is selected.)

The high dimensionality requires the use of many particles to sample the pose space sufficiently densely. Every particle comes with an increase in computational cost due to propagating the particles according to the dynamical model and the evaluation of the likelihood function. For each particle, the human body model must be rendered and compared to the extracted image descriptors. Another problem is the fact that particles tend to cluster themselves on a very small area. This is called sample impoverishment [174], and leads to a decreasing number of effective particles. Different particle sampling schemes have been proposed to overcome this problem. In [378], several common schemes are evaluated quantitatively on the task of human motion tracking.

Currently, there are two main solutions to make the problem more tractable. The first one is to use priors on the movement that can be recognized. This includes learning motion models to guide the particles more effectively, and to learn a low-dimensional space which reduces the number of particles needed. We discuss these topics in Section 2.3.3. A second solution is to spread particles more efficiently in places where a suitable local maximum is more likely. We discuss this solution below.

Sminchisescu and Triggs [330] introduce covariance scaled sampling (CSS) to guide the particles. Instead of inflating the noise component in the model of dynamics, the posterior covariance of the previous frame is inflated. Intuitively, this focuses the particles in the regions where there is uncertainty, for example due to depth ambiguities as observed in monocular tracking. In the unconstrained case and given monocular data and known segment lengths, each joint has a two-fold ambiguity. The connected limb is either placed forwards or backwards. This also means that there are two local maxima in the likelihood surface. When tracking fails, this is most likely due to choosing the wrong maximum. In [331], these ambiguities are enumerated in a tree, and the particles are allowed to 'jump' in the pose space accordingly. Deutscher *et al*. [66] introduce a different approach to guide the particles. They use simulated annealing to focus the particles on the global maxima of the posterior, at the price of multiple iterations per frame. Particles are distributed widely at initialization, and their range of movement is decreased gradually over time. Lu [205] additionally relaxes the particle fitness function at the higher levels to avoid getting trapped in local maxima.

MacCormick and Blake [208] partition the pose space into a number of lower-dimensional subspaces. Because independence between the spaces is assumed, this idea is similar to search-space decomposition. Husz *et al*. [141] observe that partitioning the pose space is difficult due to the high correlation between different body parts. They introduce a hierarchical version of partitioned sampling, and use a switching model for dynamics. Bandouch *et al*. [19] demonstrate that the strengths of annealed particle filtering and partitioned sampling are complementary with respect to initial estimates and dimensionality, and introduce a combined filtering scheme. Along the

same lines, Fontmarty *et al*. [96] combine partitioned annealed particle filtering with an importance sampling stage in order to enable automatic initialization.

2.3.2.2 Batch approaches

In a batch, or smoothing, approach, poses are optimized over a sequence of frames, instead of online. The need of propagating multiple hypotheses is not required as the globally optimal sequence of poses can be determined automatically. Plänkers and Fua [270] and Liebowitz and Carlsson [197] use least-squares minimization, Brand [40] and Navaratnam *et al*. [236] use the Viterbi algorithm to find the most probable state sequence in a hidden Markov model (HMM). Zhao and Nevatia [409] present a tracking-by-detection framework in a more advanced graphical model. State transitions for several locomotion styles are learned from motion capture data. Again, the optimal sequence is found using the Viterbi algorithm. As each state corresponds to a pose template, post-processing is used to smooth the results. Peursum *et al*. [269] investigate the effect of smoothing over filtering. A standard particle filter, an annealed particle filter, and a standard particle filter with learned motion dynamics are evaluated. No significant improvement was observed, which was attributed to the high dimensionality of the search space.

2.3.3 Motion priors

Although the human body can perform a very broad variety of movements, the set of typically performed movements is usually much smaller. Motion models can aid in performing more stable tracking, especially when only a single class of movements (e.g. walking, swimming) is regarded. However, this comes at the cost of putting a strong restriction on the poses that can be recovered.

Many prior models are derived from training data. A possible weakness of these motion models is that the ability to accurately represent the space of realizable human movements depends largely on the available training data. Therefore, the set of examples must be sufficiently large and account for the variations that can be observed while tracking the movement. We identify two main classes of motion priors. The first uses an explicit motion model to guide the tracking. The second class learns a low-dimensional activity manifold, in which tracking occurs.

2.3.3.1 Using motion models

In a tracking approach, the prediction in the next frame can be obtained by extrapolating the joint angles or joint positions given the previous frame. Such extrapolations can be linear or take into account the acceleration of the body part. However, many activities show a clear movement pattern and a specific motion model can be used to obtain accurate predictions. Most statistical motion models can only be used for specific movements, such as walking [129; 295], dancing [287], playing golf [361] or tennis [336].

Sidenbladh *et al*. [319] retrieve motion examples similar to the motion being tracked from a database. The dynamics of the example are used to propagate the

particles in a particle filter framework. Fathi and Mori [87] use the same concept, but select motion examples based on flow features of subsequent frames.

Ning *et al*. [243] constrain the propagation of the particles using physical motion constraints which are learned probabilities conditioned on the parent joint. Instead of learning models from data, human motion can be described using physical models. Rosenhahn *et al*. [303] introduce constraints to avoid that the feet intersect the ground plane. Additional constraints that originate from interacting with the environment, which are common in sport motion analysis, are modeled in [302]. Brubaker *et al*. [43] model the hips and knees as a mass-spring system. This allows them to model balance and ground contact while being able to deal with variations in walking style, mass and speed. In Brubaker and Fleet [42], the model is adapted to include the torso and ankles, which allows to recover walking movements on slopes. Vondrak *et al*. [368] regard the whole body and model each body part as a rigid object with known mass, inertial properties and geometry. As such, ground contact and interactions with objects whose locations and geometry are known can be modeled. To reduce the computational complexity due to the high search space, an example-based approach is adopted. Using k-nearest neighbor (k-NN), the closest k motion examples are selected and a weighted interpolation gives the initial pose prediction. Fossati and Fua [100] guide tracking by observing that the orientation of the person should be in the direction of the movement, and vice versa.

Pavlović *et al*. [265] use a switching linear dynamical model. Each state of the model corresponds to a particular class of poses, and the dynamics within this class are assumed linear. The work of [44] does not only model the short-term dynamics but also takes into account the history using variable length Markov models (VLMM). Clusters of elementary motion are learned from training data and clustered. State transitions in the VLMM correspond to one of the clusters. Particles are propagated according to the dynamics of the selected cluster with additional noise sampled from the covariance of the cluster. This is similar in spirit to CSS [330]. Peursum *et al*. [268] introduce a factored-state hierarchical HMM (FS-HHMM) which is similar in concept, but is more robust against noisy observations.

Wang *et al*. [379] address a slightly different problem termed motion alignment. The idea is to align 2D observations to prerecorded 3D sequences in both space and time. The work can be used to find deviations from a optimal sport motion, which acts as a very strong motion 'prior'.

2.3.3.2 Dimensionality reduction

For a given action class, the movement of individual joints is often highly correlated. Hence, a lower-dimensional latent space can be learned that still faithfully describes the possible variation in the movement [116]. Such a low-dimensional manifold is usually 2-4 dimensional, which is significantly lower than the original pose dimensionality. Tracking in such a manifold results in lower numbers of required particles and a reduced risk of getting trapped in local maxima. Manifolds are often learned for specific activities, such as walking. Despite recent work that takes into account various locomotion styles [150; 360], it remains to be researched how this can be extended to broader classes of movement.

For a model-based approach, tracking in a low-dimensional manifold requires three components. First, a mapping between original pose space to low-dimensional manifold must be learned. Second, an inverse mapping must be defined to obtain full-pose estimates for the generation of model projections. Third, it must be defined how tracking within the low-dimensional space occurs.

Principal component analysis (PCA) is a common linear dimensionality reduction technique. It has been used by Urtasun *et al*. [358] and Sidenbladh *et al*. [318], both using local optimization to avoid maintaining multiple hypotheses. Agarwal and Triggs [2] first cluster training samples with similar dynamics before applying PCA to reduce the dimensionality of each cluster. They learning a local linear auto-regression model to propagate dynamics. A class inference algorithm is able to estimate the current motion cluster and allows for smooth transitions between clusters.

Since the mapping between the original pose space and latent space is in general non-linear, linear PCA is inadequate. Algorithms such as locally linear embedding (LLE) and Isomap can learn this non-linear mapping but are not invertible. This inverse mapping is needed because the full pose representation is required for evaluation of the likelihood function. Sminchisescu and Jepson [326] use spectral embedding to learn the embedding, which is modeled as a Gaussian mixture model. Radial basis functions (RBF) are learned for the inverse mapping.

Gaussian process latent variable models (GPLVM, [189]) and locally linear coordination (LLC, [343]) do provide the inverse mapping from latent space to pose space. Li *et al*. [196] learn a mixture of factor analyzers, each of which is modeled using LLC. A simple multiple hypothesis tracker is used for tracking in the latent space.

Urtasun *et al*. [361] use a GPLVM to learn prior models for 3D human tracking. GPLVMs generate smooth mappings between pose space and latent space. They demonstrated that, for actions with limited variation, a deterministic tracking approach was sufficient to accurately recover the motion. Instead, Tian *et al*. [347] used particle filtering in the latent space to estimate 2D upper-body poses using a GPLVM. Since the GPLVM only learns a mapping from the latent space to the pose space, local distances in the pose space are not preserved. This issue is solved in the back-constrained GPLVM (BC-GPLVM), used by Hou *et al*. [131], where an additional mapping from pose to latent space is learned. Due to the distance preservation in both directions, the resulting mapping can be regarded as one-to-one. Similarly, Urtasun *et al*. [362] use a locally-linear GPLVM (LL-GPLVM) and additionally introduce sparse Gaussian processes to be able to deal with larger training sets.

The GPLVM does not model a prior over the latent space, which does not allow it to penalize drifts from the manifold of common configurations. Kanaujia *et al*. [165] combine spectral embeddings and parametric latent variable models into a sparse spectral latent variable model (SLVM) to address this issue. The learned mappings between latent and pose space are bi-directional. Their approach is used in a discriminative fashion, where a mixture of experts models the distribution in latent space, given an image observation.

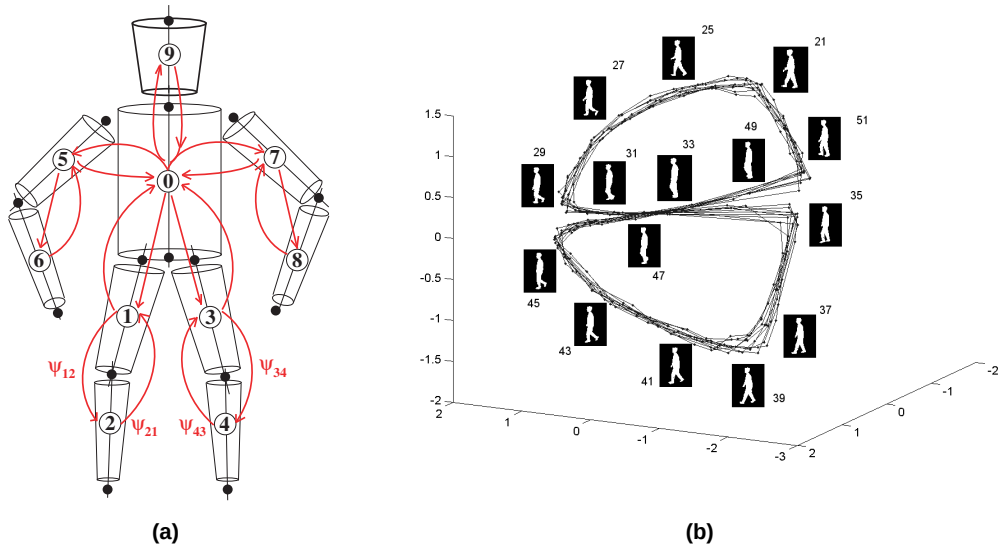
Instead of learning the embedding space from static examples and subsequently learning the dynamical model, these two tasks can be coupled. The temporal pose data not only ensures efficient tracking in the embedded space, but also regulates

the construction of the embedding space. Li *et al*. [195] simultaneously learn the low-dimensional manifold and the dynamical model. The manifold is approximated by piece-wise linear regions, and a linear dynamical model within each region is assumed. Gaussian process dynamical model (GPDM, [371]), an extension of the GPLVM, allows to model non-linear motion. The GPDM has been shown to accurately recover motion, even when learned from few training sequences [359]. Pang *et al*. [260] present the Gaussian process spatio-temporal variable model (GP-STVM), which is similar in concept, but explicitly model the spatial relationships in the pose data when learning the embedding.

The above works learn a low-dimensional manifold of the pose space. Such an approach is useful for model-based approaches, where tracking can be performed in the latent space and the full pose representation is used to generate the projection of the model. For model-free approaches, the aim is to estimate the low-dimensional representation directly given an observation. To this end, the embedding can be learned in the image space instead of the pose space. Elgammal and Lee [79] take this approach and learn the embedding using LLE. A mapping from embedding space to pose space is modeled using RBF but the approach is unable to solve ambiguities that arise from the use of silhouettes (see Fig. 2.3(b)). Wu *et al*. [396] also learn an embedding in image space, but use temporal neighbor-preserving embedding (TNPE). The mapping from latent space to pose space is approximated using a Bayesian mixture of experts, which is able to cope with observation ambiguities.

When modeling the pose space as a latent space, the mappings from pose space to image space may be unnecessarily complex. The same is true for modeling a low-dimensional representation of the image space alone. To overcome this issue, Tangkuampien and Suter [340] learn a low-dimensional representation of both spaces, and use LLE as a mapping from latent image space to latent pose space. Due to the lower dimensions, the complexity of the mapping is reduced as well. Jaeggli *et al*. [150] also model both embeddings individually, but learn sparse kernel regressors to model the mapping from latent pose space to latent image space. The approach is thus generative, and accounts for the multi-modality of the image-pose mapping. The authors further use an activity-switching mechanism and learn the embeddings and mappings for each activity individually.

Ek *et al*. [75] model the embedding in the joint space and learn mappings to both the pose and image space. By considering a back-constrained GPLVM, the mapping between pose and latent space can be considered one-to-one, whereas the multi-modal relation between latent space and image space is modeled using a mixture of regressors. Navaratnam *et al*. [235] take a similar approach but additionally incorporate unlabeled data. Elgammal and Lee [80] use a slightly different approach by learning the joint embedding in a supervised manner. They use a torus-shaped manifold to encode pose, viewpoint and shape style, and learn mappings to the pose and image space. Due to the absence of a mapping from image space to latent space in these works, initialization is still difficult and is performed using local optimization. By considering a joint latent space, it is assumed that all variance is explained in the embedded space. In many practical problems, much of the variance is specific to either the pose or the image space. Ek *et al*. [74] address this issue by decomposing the la-



(a) (b)

Figure 2.3: (**a**) Relation between body parts described in a graphical model [325] (© MIT Press, 2003) (**b**) View-based manifold for walking activity [77] (© IEEE, 2004))

tent space in a shared latent space and private latent spaces for both pose and image space. This leads to a formulation where inference can be performed efficiently in the shared latent space, whereas potential ambiguities from the image observation lie in the latent pose space.

Gupta *et al.* [119] take advantage of the strengths of both a generative and a discriminative approach. They use a GPLVM and add a mapping from image space to latent space, which is modeled as a mixture of experts. To be able to deal with activities that have various distinct performance styles (e.g. sitting motions depending on the height of a chair/ground), different mappings are learned for different context settings.

Instead of learning an arbitrary dimensionality reduction from training data, Xu *et al.* [398] only use the dimensions that correspond to the joints of either the left or the right side. Such an approach works well for activities where there is a high correlation between both sides, such as in walking. A regression function is used to obtain the joint estimates for the dimensions that are omitted.

2.3.4 3D pose recovery from 2D points

When only 2D points over a sequence of images are known, 3D poses can be estimated if a human body model is taken into account. Liebowitz and Carlsson [197] reconstruct 3D poses from 2D point correspondences from multiple views and known body segment lengths. Linear geometric reconstruction is used to recover the poses of an entire motion sequence in a batch fashion. Taylor [342] uses only a single view and recovers the entire set of pose solutions by considering the foreshortening of the segments of the model in the image. A scaled orthographic projection is assumed, which limits the approach to far views. A global scaling parameter and depth ordering of body parts must be specified manually. Since many depth orderings are

kinematically infeasible, Lee and Chen [191] construct an interpretation tree with all ambiguities that arise from forward-backward flipping and apply constraints to prune impossible configurations. Additionally, DiFranco *et al*. [68] use user-specified 3D key frames. A maximum *a posteriori* trajectory is calculated using a non-linear least squares framework, taking into account joint angle limits and smooth dynamics. Kuo *et al*. [181] focus on human locomotion and first find keyframes where the joints of the shoulders, neck and hips reside in the same plane. Camera calibration is performed on these frames. A pinhole camera model is assumed, which is often more suitable than a scaled orthographic projection. For each frame, a set of 3D pose proposals is generated. A final pose is chosen based on constraints and a learned motion model. In [261], no camera model is assumed but fixed segment ratios are used. Shen and Foroosh [314] use templates that consist of annotated 2D body point triplets. Pose transitions over time are considered, seen from different viewpoints and possible from different camera models. They use the fact that the homography that arises from a moving triplet of body points reduces to a homology when observing the same movement. As such, they select the corresponding template with associated 3D pose representation.

Howe *et al*. [137] use snippets of motion from a database to recover 3D motion given 2D points. From a sequence of 2D poses, the 3D motion is reconstructed by finding the MAP estimate of the sequence of snippets. Sigal and Black [324] use regression to obtain 3D pose representations after recovering the 2D poses using a model-based approach.

2.4 Model-free approaches

If no explicit human body model is available, a direct relation between image observation and pose must be established. In practice, this means that the image representation must generalize over variations in body dimensions, appearance and clothing. In general, a far-off view is assumed, where perspective effects are negligible. When multiple cameras are employed, calibration is assumed.

The training data must account for those parameters that we wish to recover, usually the pose representation and the viewpoint. Not all kinematically possible poses are also likely, and the training data implicitly forms a manifold in pose space. Due to the high non-linearity of this manifold, the pose space should be covered densely to obtain faithful mappings. Dimensionality reduction can be used in pose or image space to facilitate the learning of the mapping as discussed in Section 2.3.3.2.

Two main classes of pose estimation approach can be identified: example-based (Section 2.4.1) and learning-based (Section 2.4.2). Example-based approaches retain all image-pose training examples. For a given input image, a similarity search is performed and candidate poses are interpolated to obtain the pose estimate. Learning-based approaches avoid having to store a large amount of examples and approximate the mapping from image to pose space functionally by training on image-pose pairs.

Model-free algorithms automatically perform (re)initialization and can be used to initialize model-based approaches. We discuss this approach, and other combinations of model-based and model-free approaches in Section 2.4.3.

2.4.1 Example-based

Example-based approaches use a database of examples that describe poses in both image space and pose space. While no mapping from image to pose space has to be learned, the drawback of example-based approaches is the large amount of space that is needed to store the database. Moreover, matching can be computationally costly, depending on the search scheme that is used.

In its simplest form, example-based approaches encode both the image part of the database and the observation into image representations, and perform a linear search to obtain the closest matches, the nearest neighbors. The associated poses of these matches can be interpolated to allow for a more continuous range of pose estimates. Poppe [272] (see also Chapter 3) uses histogram of oriented gradient (HOG) representations, which encode edges while allowing for small variations in spatial arrangement. Results are presented from monocular and multi-view settings. In the multi-view case, the camera arrangement in training and test conditions is required to be the same. In contrast, Izo and Grimson [148] recover the location of the cameras and the walking phase of a person simultaneously in a multi-view setting. First, they estimate the gait phase from multiple cameras, and subsequently they recover the view that corresponds to each of the cameras. Their approach requires explicit training examples from all expected views, which can be prohibitive, especially when more movement classes are considered. Silhouettes, described using turning angle and Chamfer distance are considered by Howe [132]. In later work [133], optical flow information is used in addition. Fathi and Mori [87] only use motion information, which is invariant to illumination and texture.

When multiple synchronized cameras are available, a visual hull can be constructed. Van den Bergh *et al*. [25] approximate this hull using 3D haarlets, an extension to 3D of the haarlets proposed in [365]. They focus on pose recognition and learn a discriminative set of haarlets to maximize recognition performance.

Instead of using a direct distance measure, Sullivan and Carlsson [336] use deformation cost between examples and an input image. To improve the robustness of the point transferral, the spatial relationship of the body points and color information is exploited. Mori and Malik [225] employ shape contexts to encode edges. In an estimation step, the stored example are deformed to match the image observation. In this deformation, the location of the hand-labeled 2D locations of joints also changes. The most likely 2D joint estimate is found by enforcing 2D image distance consistency between body parts.

Temporal information can be used to overcome ambiguities from the image to some extent. Toyama and Blake [348] incorporate examples in a probabilistic temporal framework. By employing an HMM, they approximate a low-dimensional manifold by linear segments. Similar in concept is the work by Ong *et al*. [255], who cluster the examples and determine flow vectors for each cluster. A particle filter framework is used where the particles are guided by the flow vectors. Particle likelihoods are based on the matching distance to the closest example in the cluster.

The computational complexity of a naive nearest neighbor search is linear in the number of examples. For recovering more unconstrained movements or high number of DOF, the number of required examples grows substantially. Therefore,

Shakhnarovich *et al*. [310] introduce parameter sensitive hashing (PSH) to rapidly estimate the pose given a new image. Because of the ambiguity in the use of silhouettes alone, they use edge direction histograms within a contour. PSH is also applied in [287], where local binary silhouette features from three views are used instead.

An alternative approach to reduce the computational complexity of the matching is by storing the examples in a tree (e.g. [109]). Given an input image, a top-down matching procedure is used. Starting from the highest level node, a matching is performed for each of the child nodes. Only those subtrees that satisfy a certain criterium (e.g. threshold or best match) are further evaluated. This significantly reduces the computation time needed to select similar examples. This approach has also been taken by Yang and Lee [403], who construct the pose estimate as a linear combination of the selected examples from the bottom level of the tree. Rogez *et al*. [294] use a collection of trees. The nodes in each tree are trained to be discriminative and take into account a single dimension from a HOG representation. By using a collection of trees, many features can be used and the resulting algorithm is more robust to noise.

2.4.2 Learning-based

Learning-based approaches approximate the mapping from image to pose space functionally. The advantage of these regression methods is that inference can be performed efficiently, and training data can be discarded after training. The drawback is learning the mapping. Especially for large amounts of training examples, computation requirements might be prohibitively large.

Xu and Hogg [397] present one of the earliest uses of regression in human pose recovery. A neural network is employed to map silhouette representations to pose representations. Agarwal and Triggs [4] use non-linear relevance vector machine (RVM) regression over both linear and kernel bases to model the relation between histograms of shape contexts and 3D poses. Ambiguities are resolved using dynamics. Agarwal and Triggs [3] use direct regression to recover upper-body poses. Non-negative matrix factorization (NMF) on grid-based edge histograms is used to obtain a set of basis vectors that correspond to local features on the human body, and ignore the presence of clutter. This enables them to recover poses without relying on background segmentation. The work by Onishi *et al*. [256] is similar in spirit, and extends [272] with a noise-reduction step. Instead of applying NMF, they perform PCA in each block of cells in the grid to reduce the influence of backgrounds.

Instead of using a single view, information from multiple synchronized views can be combined into a voxel model. Ambiguities caused by using a single view are thus avoided. Also, a 3D voxel representation is independent of the camera setup. The approach is most suitable for controlled settings, where clean silhouettes can be obtained. Sun *et al*. [338] use an adaptations of the RVM to recover the pose from 3D shape context descriptors. When rotation normalization can be performed, such an approach can be used to learn view-independent regressors. Gond *et al*. [112] fit the voxel model in a 3D circular grid. This descriptor is rotation-normalized after recovering the orientation of the torso. The normalized feature representation is then used as an input for a sparse regressor. The approach has the advantage that significantly

less training data is required, at the cost of an additional normalization step.

It has been discussed before (see Section 2.3.3.2) that the space of common human poses is much smaller than the space of kinematically possible poses, and that these poses usually occupy a well-defined area in this high-dimensional space. This has led to the introduction of dimensionality reduction techniques. These techniques are also well-suited for learning-based approaches as they can simplify the regression functions. For example, Grauman *et al*. [115] describe a distribution over both multi-view silhouettes and 3D joint locations with a mixture of probabilistic PCA. A pose estimate is obtained from the Bayesian reconstruction given the image representation. Similar in concept is the work of Bowden *et al*. [39], who fit a non-linear point distribution model (PDM) to 2D position of head and hands, the 2D body contour and the 3D pose representation. The feature space is projected on a lower dimensional space and allows for reconstruction of the pose given an input image. Ong and Gong [254] include views from multiple cameras in the PDM and recover a pose from multi-view images. Rogez *et al*. [293] use single view and learn separate models. Temporal and spatial constraints are further used to solve pose ambiguities. This concept is similar to Brand's [40], who models a manifold of pose and velocity configurations with an HMM. Temporal ambiguities are resolved by recovering poses over an entire sequence by applying the Viterbi algorithm.

Taycher *et al*. [341] transform the continuous state estimation problem into a discrete one by dividing the state space into regions that approximate the posterior. The observation potential function of the CRF is learned off-line from a large number of examples. By focusing only on the regions where the prior state probability is significant, poses can be recovered in real time.

Due to depth ambiguities in image space, the mapping from image to pose space is multi-valued and cannot be determined with a single regressor. Therefore, mixtures of regressors have been introduced. These divide the image space into clusters, where a regressor is learned for each cluster. Rosales and Sclaroff [298] cluster the 2D pose space and learn specialized functions for each cluster from image descriptors to pose space. A neural network is used as mapping function. In [300], the work is extended to allow input from multiple cameras. The pose is estimated for each camera individually and in a subsequent step, the hypotheses are combined into a set of self-consistent 3D pose hypotheses. Thayananthan *et al*. [344] use a mixture of regressors but validate the pose estimate for each by matching it against the input image to select the most likely pose. A similar approach is used by Sminchisescu *et al*. [327], who jointly learn mappings between image and pose space. The processes are guaranteed to converge to equilibrium. During inference, the results of the mapping from image to pose is validated using the mapping back.

Sminchisescu *et al*. [329] take a probabilistic approach and model the multi-valued nature of the mapping with Bayesian mixture of experts (BME). Each expert has an associated gating function, which gives the conditional probability that the regressor should be used given an input image. Guo and Qian [118] adapt the initialization using k-means, and use stereo observations to reduce the multi-modality of the mapping. Ning *et al*. [242] initialize the experts on a partitioned subset of the image space. In the BME framework, experts and gating functions are learned

simultaneously. This requires a double-loop optimization approach, which is computationally costly. Therefore, Bo *et al*. [33] train both models sequentially, which results in a decrease of both memory and computation requirements. Their algorithm thus can handle much larger numbers of examples. Bo and Sminchisescu [32] observe that there are often correlations in the output space, in addition to correlations in the image space. They introduce Twin Gaussian Processes (TGP) to account for these correlations.

Usually, not the whole image representation is useful for learning the regression. Redundancy and noise in the training data can thus affect the learning and performance of the regressors. This can be avoided by selecting only the relevant features. Additionally, this lowers the dimensionality of the image space, and thus the complexity of the regressor. Ning *et al*. [244] jointly learn the BME regressors and the selection of visual words in a supervised manner. A similar approach is taken by Kanaujia *et al*. [164], who focus on hierarchical image representations and semi-supervised learning. Okada and Soatto [253] discriminatively select those orientations within HOG cells that are meaningful for predefined class of poses. An input image is first classified to a pose class, before recovering the pose. Bissacco *et al*. [28] use boosting to select a limited set of discriminative binary edge features and to learn the mapping directly.

Instead of learning the regression function offline, Urtasun and Darrell [357] learn it online, given an input image. With an example-based approach, the closest examples are selected. A local regression is then learned from these matches. Their approach can handle large numbers of examples, but is computationally more costly due to the selection of the nearest neighbors.

Learning these mappings depends largely on the availability of pose-observation pairs, which are difficult to obtain, especially in more unconstrained scenarios. Several authors have used synthetic observations generated by character modeling software (e.g. [4; 327]). Instead, Navaratnam *et al*. [234] use unlabeled examples to improve the regression functions. These examples can be easily obtained by using images of humans and by considering motion capture data.

2.4.3 Combined model-free and model-based

Both model-based and model-free approaches have their relative advantages and disadvantages. By using them jointly, one can combine the advantages of both, while partially overcoming the disadvantages of either.

Sigal *et al*. [320] use a mixture of regressors to give an initial distribution over the pose and shape space. A model-based approach is subsequently used to refine both pose and body shape estimates. As such, poses can be recovered more efficiently without using temporal information or the need to manually initialize. Rosales and Sclaroff [299] use the same concept, but present both a deterministic and a probabilistic inference algorithm. Gupta *et al*. [119] use a discriminative mapping in combination with a GPLVM. The mappings are conditioned on the context, such as the seat height for sitting motions.

Micilotta *et al*. [214] present an approach that works in the opposite direction. First, a 2D assembly of body parts is found. This representation is used as input to an

example-based approach to obtain a 3D pose representation, similar to [324].

A special case of combining model-free and model-based approach is the tracking-as-recognition approach. Poses are recognized using pose templates, and the results are refined over time using a generative model. For example, Fossati *et al*. [99] recognize walking poses where the feet are furthest apart using the template matching approach of Dimitrijevic *et al*. [69]. This pose is distinctive and can be recognized with high recall from a variety of viewpoints. A learned motion model, specific to walking, is used to interpolate between detections. Finally, a model-based approach is used to refine the poses. In Fossati *et al*. [98], the silhouette matching is replaced by a more robust matching which uses regions of moving edges. Ramanan *et al*. [283] also find typical poses, but construct an appearance model from it, which is further used in the model-based tracking process.

2.5 Discussion

Human motion analysis is a challenging problem due to large variations in human motion and appearance, camera viewpoint and environment settings. On the other hand, we know much about people's physical appearance and movements. The key point for successful human motion analysis is to use this knowledge effectively. Over the last two decades, a large amount of research has been conducted. Human body models that were initially described in 2D have now evolved into highly articulated 3D models. Deterministic linear tracking has been replaced by sampling-based tracking frameworks that evaluate the likelihood function efficiently. Machine learning plays an increasingly important role in human motion analysis, and will continue to do so.

For each of the methodologies described in this survey, prior knowledge about human movement or appearance is incorporated increasingly efficiently. But although many of these advances have led to impressive results given the complexity of the task, the evaluation domain is still limited. Not unusually, it is assumed that a person has been found in the image in a preprocessing step. Furthermore, assumptions about the viewpoint, environment, appearance and motion are often required.

We expect that combining methodologies is the solution to use prior knowledge even more effectively. Indeed, recent work explores these kinds of combinations. While much research is needed, these works are certainly promising. For example, model-based and model-free approaches have been combined [119; 320] to allow for automatic initialization and recovery. Another promising direction of research is the recent combination of bottom-up and top-down approaches. This has led to more efficient tracking frameworks. Model-free approaches are increasingly considering real data with cluttered backgrounds [28; 253] and occlusions (see Section 3.3), problems that have often been ignored.

In addition, the role of context should be used more explicitly. Human motion analysis provides input for reasoning about actions and intentions. Conversely, context can be used as input for human motion analysis, other than implicitly by assuming a fixed domain. Recent work aims at learning models that take into account the context [44; 119; 268]. The role of human motion models, and how they generalize to broader domains remains to be investigated. Also, the suitability of low-dimensional

latent spaces for recovery of more spontaneous movement needs to be assessed.

The automatic recovery of human poses and motion has many applications, ranging from real-time human-computer interaction to precise sports motion analysis. Also, new applications are emerging. Given the large amount of available images and video on the internet, automatic detection of humans and their poses [27; 178; 239] and the subsequent recognition and interpretation of their actions are needed to allow for searching. On the other hand, the retrieval of such data should also be facilitated [91].

From a practical perspective, evaluation of motion analysis algorithms requires a common datasets, representative for a broad range of domains (indoor, static scenes, and dynamic, cluttered scenes with multiple persons). Challenging scenarios are likely to be beneficial to advance the state of the art. However, care should be taken in focussing the datasets to realistic applications such as human-computer interaction, surveillance or sports motion analysis. Each of these will have its own set of constraints and challenges. Datasets should consist of ground truth data and synchronized image sequences to be used by many different pose recovery approaches. In addition, common criteria (accuracy, smoothness, speed) for evaluation are needed. The recently introduced HumanEva database [322] is a good first step in this direction. When the evaluation criteria are generally accepted, this will contribute significantly in determining promising directions of research.

3

Example-based human pose recovery using HOGs

Discriminative human pose recovery approaches can be applied in real-time applications, but are less flexible in terms of encoding of parameters compared to generative approaches. Variations such as visual appearance and viewpoint should be encoded either explicitly in the parameter space, or implicitly in the image representation. In this research, we take a discriminative approach where we explicitly encode viewpoint, as changes in viewpoint have a large impact on the image representation. We require our image representation to implicitly encode lighting variations, and variations in body dimensions and clothing.

The focus of this chapter is on robust invariant image representations. Discriminative pose recovery approaches are either example-based, or regression-based. Regression-based approaches allow for faster evaluation but their precise implementation and parameter setting (number of experts, regression function, learning of the regression and gating functions) influence the performance. This is undesirable, since it is less intuitive to attribute the performance to image representation, or regression approach. Therefore, we use an example-based approach instead.

In Section 3.2, we present our example-based approach where we use histograms of oriented gradients (HOG, [58]), a holistic image representation. The n visually most similar examples (nearest neighbors) are selected and the final pose is estimated to be the weighted interpolation of the n corresponding poses. Our approach is evaluated on the HumanEva dataset, and we discuss previously obtained results on the same dataset. An early version of this section appeared as [272].

In realistic scenarios, partial occlusion of the human figure in the image due to other persons or objects in the environment will be common. In Section 3.3, we adapt our approach to handle partial occlusion, if these can be predicted from a foreground segmentation process.

For our example-based approach, we assume that the location and scale of a human figure can be detected from video. When occlusions occur, we also require that these areas are labelled. Due to the importance of this preliminary step, we discuss the process of human detection in slightly more detail in Section 3.1.

3.1 Preliminary: human detection

Human detection and pose recovery can be seen as complementary tasks. In the detection task, the aim is to generalize over different poses, whereas in the recovery task, one wants to discriminate between them. We advocate a separation of these tasks. This eliminates the need to perform human detection and pose recovery simultaneously, as this would require large amounts of pose-annotated training pairs, which are costly to obtain.

The detection of human subjects from images is an important first step in the analysis of human pose or action. Only recently, there has been an increased interest in this topic (see [179] for an overview). In general, human detection methods are either holistic, or part-based. A *holistic* approach considers the human body as a whole [58; 109]. In many cases, a retinoscopic representation is used, where the human is assumed to be centered within a defined region of interest (ROI), or window. Human detection is performed by sliding the window over the image, and performing binary classification at each location. Dalal and Triggs [58] train a support vector machine on positive and negative examples, encoded as HOGs. All training examples are retained in [109], where Chamfer matching between a large set of pedestrian examples is performed hierarchically. Dong *et al*. [71] explicitly take into account inter-human occlusion. They extract foreground blobs of a single person, or a group of people. An example-based approach, assuming that for each blob the corresponding number of persons with their exact locations are annotated, is used to segment each person individually. Earlier work by Elgammal and Davis [76] used known color distributions of each person, to segment persons under occlusion. The advantage of holistic approaches is that they generate relatively few false positives since much information about the human body can be incorporated. The main drawback of a holistic approach is that occlusions cannot be dealt with without closer inspection of the scene.

In contrast, *part-based* approaches divide the human body into several parts, each of which can be modeled individually. Human detection is performed by looking at assemblies of these individual parts (e.g. [218; 223; 393]). Mohan *et al*. [223] find the head, legs, and the separate arms in a window by applying component-based classifiers. Then, an SVM over the individual part detectors is used to classify the entire window as human or non-human. The work of Mikolajczyk *et al*. [218] is similar in nature, but the focus is on the face and shoulders, which are encoded for frontal and side views separately. Another part-based approach is proposed by Felzenszwalb *et al*. [90]. They describe a person with a deformable part model, where each part is discriminatively learned from HOG descriptors. Niebles *et al*. [239] use a similar approach but reduce the search space by applying a holistic detector first and relying on temporal continuity. Recent work by Wu and Nevatia [393] takes into account occlusions between persons in the scene. Such an approach is not only able to detect persons, but can also determine which part of the observation is occluded. In recent work [394], they extend their approach to output pixel-level segmentations. Lin *et al*. [199] address the problem of finding suitable assemblies by introducing a tree of parts. Using re-evaluation, their method is also able to segment occluded persons

from an image. They extend their approach in [198] to better handle variation in pose.

In general, part-based approaches generate many false positives for individual body segments. This can be explained since a body segment alone is often less discriminative compared to a full body. However, part-based approaches have a number of advantages over holistic methods. First, geometric constraints can be encoded efficiently. Second, by learning the detectors for parts individually, the combinatorial problem is effectively decomposed. Therefore, fewer training data of body-part templates is needed. Third, given a sensible assembly algorithm, humans can still be detected and segmented from the image even if parts are missing. This allows part-based methods to cope with partial occlusions from the environment or other persons.

Summarized, recent work has greatly advanced the quality of human detection. Especially the successful combination of detection and segmentation leaves us to believe that it is realistic to assume that a separation into foreground, background, and occluded area can be made. In the remainder of this chapter, we assume that such a segmentation is available. In Section 3.2, we will use descriptors without the presence of occlusions. Subsequently, in Section 3.3, we adapt our approach to deal with occlusions.

3.2 Pose recovery using histograms of oriented gradients

To describe an image we can either use a holistic descriptor or a local (or patch-based) descriptor. The former encodes the image observation as a whole. Local deformations in the image will affect the entire descriptor. In contrast, local descriptors describe the image observation as a collection of local regions. Usually, these regions are extracted at interest points (local features), which are expected to be invariant to changes in viewpoint and illumination [216; 354]. Currently, a popular local descriptor is the scale invariant feature transform (SIFT, [203]) and extensions (SIFT-PCA, [167] and GLOH, [217]). Local descriptors have the advantage that they can cope with variations in illumination, pose and viewpoint to some extent. However, they strongly rely on robust extraction of interest points, which might be difficult due to differences in subject and background appearance. Moreover, extraction of local descriptors is more time-consuming due to the localization of interest points and the calculation of the local descriptor. Also, matching of bags of local descriptors is less straight-forward. Therefore, we use a holistic descriptor in our work.

In this section, we present an example-based approach to human pose recovery. We use histograms of oriented gradients as image representation. This holistic representation has been introduced by Dalal and Triggs [58]. Their HOG descriptor is inspired by work on orientation histograms [101], but uses dense sampling instead. The key idea is to calculate histograms of oriented gradients (edges) within each cell of a regular spatial grid. This grid has a fixed number of cells, which cover an area that is determined by a rectangular region of interest (ROI). The HOG descriptor is a concatenation of all cell histograms. Several alternatives to HOGs have been proposed in literature. Levi and Weiss [194] use edge orientation histograms that contain ratios

between orientation responses, dominant orientation and symmetry features, calculated exhaustively over all rectangular subwindows of an image. Adaboost is used to select the relevant features. In contrast, HOGs use a fixed spatial structure, which allows for direct matching. The pyramid of histograms of oriented gradients (PHOG), proposed by Bosch *et al*. [37] is a generalization of the HOG where the notion of a block of cells is extended to multiple levels. At the lowest level, the ROI is described as a single edge orientation histogram. For each higher pyramid level, a division into 2×2 cells is made. The PHOG approach is suitable when there is variation in the localization of the ROI but restricts the number of rows and columns in the grid to be equal and to be powers of 2. As the height of a human figure in the image is larger than its width, we use the original HOG concept.

HOGs have been used for several human motion analysis tasks. Dalal and Triggs initially used HOGs for pedestrian detection, a binary classification task. Variations in clothing, lighting, body dimensions, but also viewpoint and pose, were implicitly encoded. Such an approach is reasonable since there are clear cues such as head and shoulder lines, which remain present also when seen from different viewpoints. Gandhi and Trivedi [107] use HOG descriptors to classify the orientation of pedestrians, thus explicitly encoding the (relative) viewpoint. Thurau [345] used HOGs to model human shape for human action recognition. Both [202] and [46] use body part classifiers based on HOG descriptors. Such an approach is suitable for 2D location of limbs. However, without strong pose priors (such as used in [276]), lifting these to 3D will lead to ambiguities as there is no verification step where the observation is used in a holistic manner.

While HOGs have been shown to be robust descriptors for the aforementioned tasks, we believe that HOG descriptors are even sufficiently rich for recovery of human poses, including the viewpoint. This task is, however, more demanding as we do not have to distinguish between a small number of classes, but instead aim at *regression* of 60-dimensional poses. Moreover, the HOG descriptors still have to be invariant to lighting, clothing and body dimensions.

Since we need to recover more information from the HOG descriptors, we also require more precise HOG extraction. The above mentioned works extract the HOG descriptors directly from the image, which has two drawbacks. First, the ROI needs to be determined, which is computationally expensive. The ROI can vary in position and scale (we do not regard rotation, upright recordings are assumed), and many possible ROI candidates within an image have to be validated. Zhu *et al*. [414] introduce an efficient approach based on the integral image [278], but real-time performance still cannot achieved. Moreover, there will be false positives in the neighborhood of the actual ROI, which makes determination of the actual location and scale difficult.

Second, there is the problem of background clutter. Edges within the ROI that do not belong to the person, but to the background, will affect the HOG descriptor. This is undesirable as it makes reliable pose recovery dependent on even backgrounds, which greatly limits generalization. Therefore, a number of works have explored ways to learn which edges belong to the foreground. Agarwal and Triggs [3] use non-negative matrix factorization to suppress background edges. They demonstrate their work on recovery of frontal poses. Both Sminchisescu *et al*. [327], and Okada

and Soatto [253] implicitly determine a set of discriminative features by learning regression functions from the HOG-space to the pose space. Due to the use of multiple regressors, this selection is pose-dependent. Bissacco and Soatto [27] perform pose recognition using latent Dirichlet allocation. This is a generative model that describes the distribution of HOG features conditioned on a latent variable, which corresponds to part of the pose space. Again, only those edge orientations that correspond to the foreground are used in the modeling.

In many cases, silhouettes can be obtained relatively reliably using background substraction. We assume that such a segmentation into foreground and background can be made. Employing this segmentation has two main advantages. First, determination of the ROI is straightforward. This relieves the burden on the detection task, as only a single detection window has to be processed. This will significantly aid in achieving real-time performance. Second, by considering only foreground edges, we effectively ignore background clutter. The resulting HOG descriptor is therefore not dependent on the background, which increases generalization. In this section, we explain the steps of our approach, and show that poses can be recovered accurately, even when foreground segmentation is noisy.

In our contribution, we do not regard the temporal aspect, nor do we apply any measures to reduce the computational complexity. This allows us to focus on the performance of the HOG descriptors. In Section 3.2.1, we discuss our HOG variant, and how we obtain the descriptor from an image. The nearest neighbor pose recovery approach is explained in Section 3.2.2. Our experiments on the HumanEva datasets are presented in Section 3.2.3 and 3.2.4. A discussion of our results, and a comparison with previous work appears in Section 3.2.5.

3.2.1 Histogram of oriented gradients

Dalal and Triggs [58] proposed histograms of oriented gradients as an image descriptor to localize pedestrians in cluttered images. The motivation for the use of gradients is that they are to some extent invariant to lighting changes. HOGs additionally preserve spatial ordering to some extend, which has been found to be of key importance for effective human pose recovery [275].

Our descriptor differs from the HOG descriptor as described by Dalal and Triggs. First, we only take into account the edges within a foreground mask. This requires background segmentation, but allows us to focus only on those edges that are meaningful. Second, we don't use the notion of (overlapping) blocks, which results in a significantly reduced descriptor size. Third, we do not apply color normalization, which further reduces the computational costs of calculating the descriptor. Fourth, we use a different grid size. In our main experiment (Section 3.2.3), we divide the ROI into a grid with 6 rows and 5 columns. This is an arbitrary choice, but the height of each cell roughly corresponds with the height of the head in a standing position. Similarly, in a relaxed standing pose, the body covers approximately 3 columns horizontally. We call this setting HOG-*F*-5×6. The *F* stands for *fit*, as the ROI is the minimum enclosing rectangle of the foreground region, which exactly fits the observation. In Section 3.2.4.1, we experiment with different grid sizes and ROI settings.

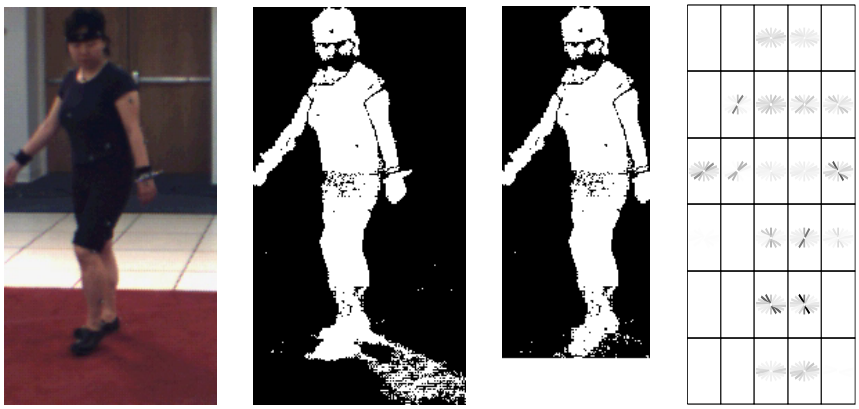
Within each cell in the grid, we calculate the orientation and magnitude of each

pixel that appears in the foreground mask. We apply a $[-1\ 0\ 1]$ gradient filter to each pixel in horizontal and vertical direction independently. The orientation of the edge is given by the angle between these two derivatives. The magnitude is given by the square root of the sum of the two squared derivatives. We divide the absolute orientations over 9 equally sized bins in the 0°- 180° range. Each pixel contributes the magnitude of its orientation to the according histogram bin, which results in a 9-bin histogram per cell. Note that this binning is slightly different than proposed in [58], where votes are interpolated bi-linearly between the neighboring cells and orientation bins. The total length of the descriptor is 270. The entire descriptor is normalized to unit length to overcome differences in scale. This normalization makes the descriptor holistic, as local variations affect the entire descriptor. To further reduce the size of the HOG descriptor, PCA is applied by Lu and Little [204] and Onishi *et al*. [256] in the context of human motion analysis. While increased robustness to background noise and illumination is reported, we did not apply PCA and used the full descriptors instead.

3.2.1.1 Determination of ROI and foreground mask

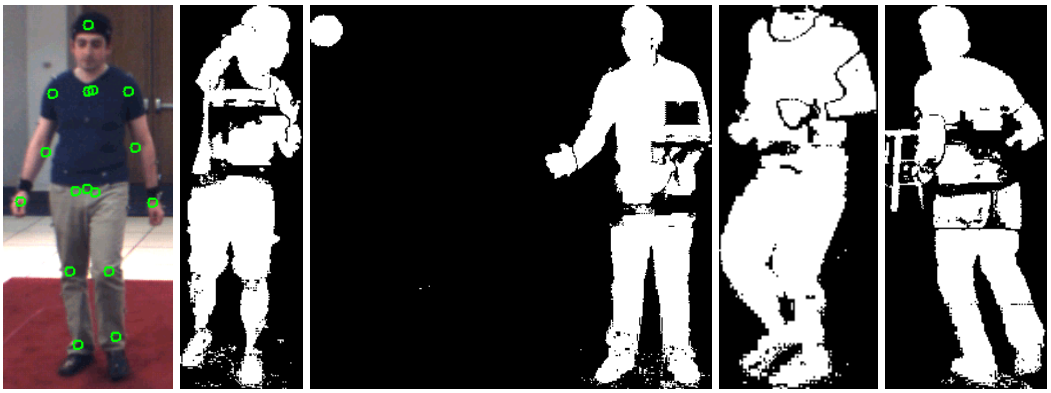
We calculate HOGs within an image's region of interest (ROI), in our case the bounding box around the subject. While HOGs can be used to determine this region, as in [58; 345; 414], we rely on background subtraction. As discussed previously, this significantly speeds up the process, and we suppress background edges at the same time. We describe the process here in detail to allow for replication. First, we apply the background subtraction with the suggested risk values, as included in the HumanEva source code [322]. The background is modeled as a mixture model with 3 Gaussians per color channel. The minimum enclosing box of all foreground areas larger than 600 pixels is obtained. After conversion to HSV color space, we apply shadow removal in the lower 20% of the ROI. Pixels that have a saturation that is between 0 and 25 higher than the saturation of any of the means in the background mixture model, are removed from the foreground mask. We again obtain the minimum enclosing box, which is our final ROI. Figure 3.1 shows an example of background subtraction and shadow removal, and displays the HOG descriptor.

It may seem that our approach is highly sensitive to good background subtraction, but the shadow removal is only needed to ensure that the ROI fits the subject reasonably. For certain cases, we slightly adjusted the parameters. For camera 1 in HumanEva-I, only for subject 2, we multiplied the risk with factor 10^{12} to remove artefacts from the foreground. For cameras 2 and 3 in HumanEva-I, we lowered the shadow threshold to 10. We did not use the additional four grayscale cameras in HumanEva-I. For HumanEva-II, we reduced the background risk with factor 10^{50}, only for camera 3. Still, errors in the background segmentation frequently result in incorrect determination of the ROI and inaccurate foreground masks (see also Figure 3.2(b-e)).



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

Figure 3.1: Example of foreground mask calculation: (**a**) original image, cropped to fit the initial ROI, (**b**) the foreground region, and (**c**) shadow removed. (**d**) shows the HOG-*F*-5×6 descriptor. Note the higher histogram values (darker lines) for the dominant direction of the legs and arms.



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

Figure 3.2: (**a**) Locations of the 20 joints. (**b-e**) Background subtraction errors that result in inaccurate foreground masks and incorrect placement of the ROI.

3.2.2 Pose recovery using nearest neighbor interpolation

In an example-based approach, each image observation is encoded and matched against an example set of encoded observations. We use the previously described HOGs as encodings. To match a HOG with those in the example set, we need to define a distance measure between the two descriptors. We performed a small-scale experiment with Manhattan, Euclidian, cosine and χ^2 distance. Consistent with earlier findings, Manhattan and Euclidian distance proved to be most suitable. In this work, our focus is on real-time performance. Therefore, we use Manhattan distance since it has a lower computational cost.

Matching a HOG with the entire example set results in a distance value for each of the m examples. We could choose the example with the lowest distance, as this is

the example that best matches the image observation. However, in practice, taking the n best matches (nearest neighbors) results in more accurate pose recovery. It should be noted that n is the only parameter in our approach. Of course, n will depend on the number of examples in the example set that are close to the presented frame. Here we use $n = 25$, in accordance with [275]. To determine the final pose estimate, we use the poses that correspond to the n best examples. We determine the final pose estimate $\boldsymbol{p}$, the normalized weighted interpolation of these poses, as $\boldsymbol{p} = \sum_{i=1..n} w^i \boldsymbol{p^i} / (\sum_{j=1..n} w^j + \delta)$ and $w^i = \frac{1}{d^i + \delta}$ $(1 \leq i \leq n)$. Here, δ is a small number to avoid division by zero in the rare case that retrieved examples exactly match. $\boldsymbol{p^i}$, d^i and w^i correspond to the pose vector, HOG distance value and weight of the i^{th} best matching example, respectively. This implies that close HOG matches contribute more to the final estimate than HOG matches at a larger distance. One word of caution is in its place here. Since we interpolate poses, the final joint estimates are likely to lie closer to the mean distance for this joint, so closer to the body. This effect is especially visible for examples that have similar image observations but are distant in pose space.

Since we do not determine the correspondences between our localized subject in the image and the estimated pose, we are not able to estimate the global position of each joint. Instead, we report the distances of each joint relative to the pelvis (*torsoDistal*) joint.

The temporal aspect of human movement can be used to improve the accuracy of an example-based approach. Howe [136] recovers poses over an entire sequence (batch). This ensures consistency in pose over time, which is particularly useful when only a single view is used. Ong *et al.* [255] use a tracking approach, where a dynamical model puts a prior on the poses in the next time frame. This guarantees temporal consistency, and reduces the number of evaluations needed since only examples with a non-negligible prior have to be evaluated. In this work, we do not enforce temporal consistency. This would require us to apply a dynamical model, which might be too restrictive. Also, such an approach is dependent on the frame rate of the recorded sequence. Other measures to reduce the order of example-based algorithms to sub-linear include hashing [151; 310]. Also, a hierarchical organization of the examples can reduce the complexity of the algorithm [109; 402]. Such measures have proven to significantly reduce computation time, while increasing the recovery error only slightly. Since we focus here on the performance of the image representation, we do not apply any measures to reduce the search time.

3.2.3 Experiment results

In this section, we describe the HumanEva dataset that we have used for the evaluation of our approach. The construction of the example sets is described in Section 3.2.3.2. Results for HumanEva-I and HumanEva-II, the two available parts of the dataset, are described in Sections 3.2.3.3 and 3.2.3.4, respectively. We describe additional experiments in Section 3.2.4.

3.2.3.1 HumanEva dataset

The HumanEva dataset [322] consists of two parts: HumanEva-I and HumanEva-II. HumanEva-I contains several sequences, divided into training, validation and test sets. HumanEva-II consists of two test sequences.

The training and validation sequences in HumanEva-I contain synchronized video and motion capture (mocap) data. There are 4 subjects that perform 5 actions (Walking, Jog, Box, Gesture, Throw/Catch). In addition, there is one sequence for each subject-action pair that contains only mocap data. In the test set of HumanEva-I, all these subject-action pairs also appear. Also, for each subject, there is an additional Combo sequence that contains walking, jogging movements, and some additional balancing movements that do not appear in any training or validation trial. The two test sequences of HumanEva-II also contain Combo movements, performed by subjects 2 and 4. Example frames of HumanEva-I and HumanEva-II are shown in Figures 3.3 and 3.7, respectively.

The walking and jogging actions are performed by moving in a circle, counterclockwise. Boxing, gesturing, and throwing and catching a ball are performed while facing camera 1. There is some variation in these sequences, though. Especially, the body orientation while catching the ball is heavily dependant on were the ball appears.

In HumanEva-I, each sequence has been recorded with 3 color cameras and 4 grayscale cameras. These have all been synchronized, and calibrated. For HumanEva-II, a different setup is used, with only 4 color cameras. The frame rate of the video sequences is 60 frames per second. The dataset comes with source code to temporally align different video streams with the motion capture data. Also, background subtraction is included, that describes the background with a Gaussian mixture model. We have used these provided algorithms where possible. In our experiments, we use only the color cameras. For the monocular case, we only regard camera 1. Sequences of subject 4 were not evaluated due to difficulties with the background subtraction.

For the test sets, ground truth pose information is held back. An online validation system, as described in [322], is used to validate the pose recovery results. This system ensures that results of different parties can be compared, and frustrates parameter tuning. Specifically, ground truth information in our case are the 3D positions of 20 key joints, relative to the pelvis (*torsoDistal*) joint (see Figure 3.2(a)). This is the full set of joints, and we report the root mean squared error (RMSE) in *mm*, averaged over all joints, as described in [322]. Evaluation of the Combo and Throw/Catch test sequences of subject 1 failed repeatedly. Consequently, we can not report our results on these trials.

3.2.3.2 Example sets

We describe our two example sets, one for monocular pose recovery (T1), and one for recovery using three cameras simultaneously (T3). For the monocular example set T1, we associate the HOGs for an individual view with their corresponding poses. Only the examples that contain valid mocap data are included in the example set.

When given a new image observation, together with the knowledge from which

Action	Subject 1	Subject 2	Subject 3	Total
Walking	1176	876	895	2947
Jog	439	795	831	2065
Throw/Catch	217	806	0	1023
Gestures	801	681	214	1696
Box	502	464	933	1899
Combo	0	0	0	0
Total	3135	3622	2873	9630

Table 3.1: Number of valid examples per action and subject in HumanEva-I training and validation sequences.

camera the observation is obtained, we can estimate the relative pose. We observe that the elevation (rotation in vertical direction) and roll (rotation around line of sight) of all cameras are approximately the same. In other words, the orientation of all cameras is almost equal except for the orientation around a vertical axis. If we would rotate the subject in the scene around a vertical axis, we would theoretically be able to generate very similar observations for all cameras. In practice, view-specific parameters such as backgrounds and lighting conditions are likely to result in observations that are somewhat different. However, we want our approach to be robust against these image deformations and therefore, we perform this rotation virtually. This has the additional advantage that the number of examples is effectively tripled, resulting in a total of 28,890.

We transform the mocap data in such a way that we obtain the joint positions as if we were looking through another camera. With an observation from camera i, and the projection onto camera j, our pose vector $\boldsymbol{p_i} = (x_i, y_i, z_i, 1)^T$ is transformed into $\boldsymbol{p_j}$ as follows: $\mathbf{p_j} = M_j M_i^{-1} \boldsymbol{p_i}$, where M_i and M_j are the rotation matrices of cameras i and j, respectively.

In T3, we combine the HOGs of the three views into a single HOG descriptor of length 810. This combined descriptor is larger, and contains the same pose seen from multiple views. We therefore expect increased pose recovery accuracy over the monocular descriptors. However, combining our HOGs comes at a cost of a reduced number of examples, compared to T1. For each frame with valid mocap, we have exactly one example. The total amount of examples m that we can obtain is 9,630. Table 3.1 summarizes the origins of the examples.

Combining all views into one descriptor has some drawbacks. When the setup of the cameras changes, the descriptor cannot be used anymore. The relative orientations of the cameras are encoded implicitly in the combined descriptors. Practically, this means that we cannot evaluate the HumanEva-II sequences with our example sets, since these are obtained from the HumanEva-I setting. Another drawback arises when one of the views contains inaccurate segmentations. This can, in some cases, render the example useless. Alternatively, we could have treated each view independently, which would be similar to using the T1 example set, but with three simultaneous observations. This would avoid the limitations of a fixed camera setup, and inaccurate segmentation is a single view would not affect the entire descriptor. However, we would then not be able to take advantage of the combined information of all

views, which allows us to disambiguate between poses.

3.2.3.3 Results for HumanEva-I

The HumanEva-I test sequences are performed with the same camera setup as the example sets. Also, the same test subjects appear in these videos. Except for action Combo, all action-subject combinations also exist as training sequences. We did not perform evaluations for subject 4, due to background segmentation errors.

For each sequence, we evaluate the performance using both T1 and T3. For T1, we use the image observations from camera 1. For T3, we use the combined image observations of all three cameras. We omitted all frames for which the mocap data was invalid from our results, but we show them in the graphs. Also, the first 5 frames of each sequence were removed since these frames were duplicated in the decoding of the video. The results, broken down by action, subject and example set, are summarized in Tables 3.2 and 3.3. Sample frames are shown in Figure 3.3. Some recovery results are presented in Figure 3.4.

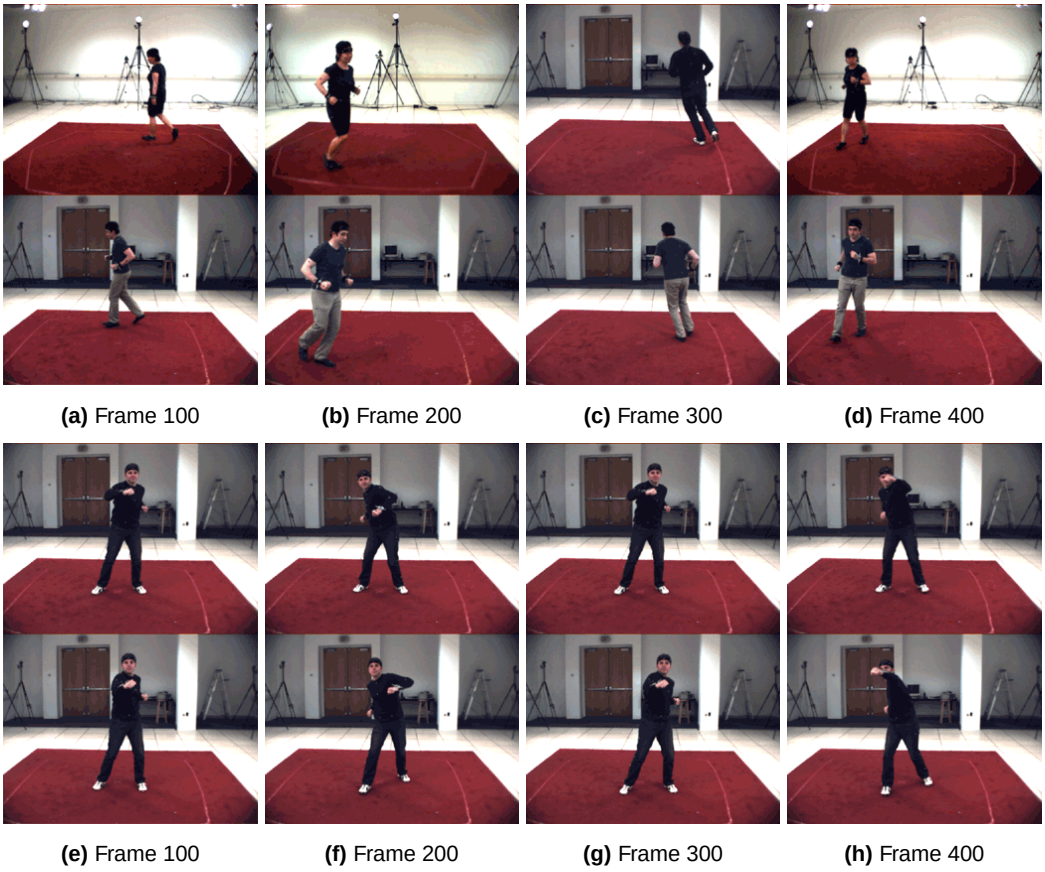
Action	Subject 1	Subject 2	Subject 3	Average
Walking	41.24 (16.84)	39.56 (26.75)	55.27 (21.70)	45.36 (21.76)
Jog	46.38 (18.60)	38.02 (9.97)	47.35 (23.47)	43.92 (17.35)
Throw/Catch		69.49 (31.94)	111.71 (33.38)	90.56 (32.66)
Gestures	26.38 (13.51)	75.13 (28.10)	75.29 (11.09)	58.93 (17.60)
Box	79.71 (27.71)	103.37 (45.08)	100.35 (52.31)	94.48 (41.70)
Combo		69.84 (49.34)	106.11 (79.74)	87.98 (64.54)
Average	48.32 (19.17)	65.90 (31.86)	82.68 (39.95)	65.63 (30.33)

Table 3.2: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-I test sequences, evaluated with a single camera (C1) using T1.

Action	Subject 1	Subject 2	Subject 3	Average
Walking	37.54 (11.95)	40.09 (23.86)	55.25 (25.07)	44.29 (20.29)
Jog	45.21 (13.74)	37.65 (12.20)	45.37 (18.32)	42.74 (14.75)
Throw/Catch		57.61 (23.75)	92.79 (31.75)	75.20 (27.75)
Gestures	23.69 (7.18)	72.83 (26.32)	56.11 (6.44)	50.88 (13.31)
Box	88.67 (36.20)	91.28 (41.19)	92.28 (47.86)	90.74 (41,75)
Combo		71.89 (52.17)	83.89 (53.10)	77.89 (52.64)
Average	48.78 (17.27)	61.89 (29.92)	70.95 (30.42)	60.54 (25.87)

Table 3.3: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-I test sequences, evaluated with all three cameras (C1–C3) using T3.

To be able to interpret our results, we first introduce a baseline. Since we interpolate the poses of the best matches, our pose estimate will always be within the convex hull of poses in the example set. Therefore, we choose our baseline to be the mean distance for randomly selecting a pose from the example set with $n = 1$. The mean distance per joint for T3 is then 299.87 *mm*. For T1, the distance is slightly lower, 291.54 *mm*. Such a baseline is reasonable since we aim at recovering poses that are



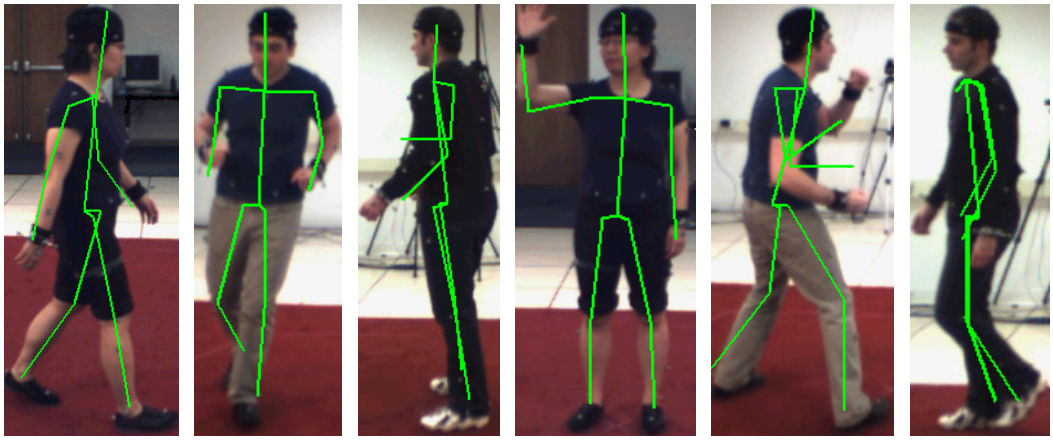
(a) Frame 100 **(b)** Frame 200 **(c)** Frame 300 **(d)** Frame 400

(e) Frame 100 **(f)** Frame 200 **(g)** Frame 300 **(h)** Frame 400

Figure 3.3: Single best estimate (top row) and original frame (bottom row) for the HumanEva-I Jog sequence performed by subject 2 (**a-d**) and HumanEva-I Throw/Catch sequence performed by subject 3 (**e-h**), both evaluated using a single camera (C1).

in the convex hull of the example set. For actions not included in the example set, such as the Combo action, errors might be much higher than this baseline.

The first thing to notice is the relatively small difference between the performance of our monocular tests, and those using three cameras. We obtained mean errors over all subjects and sequences of 65.63 and 60.54 *mm*, respectively. This observation seems to contradict earlier work by Grauman *et al.* [115], who use concatenated descriptors of silhouettes. They found that pose recovery accuracy increases with the number of views that is used. However, there are two main differences between their approach and ours that could explain the discrepancy. First, we use HOG descriptors that encode edge information. These are helpful when differentiating between front and back poses, thus avoid depth ambiguities. With silhouettes from a single view, such ambiguities cannot be resolved. Second, Grauman *et al.* [115] use approximate Earth-Movers distance, which is more computationally demanding but might overcome small errors due to segmentation. Since, in the multi-view setting, we use the three observations as a single descriptor, segmentation errors in each of the views introduce noise in the entire descriptor. This is true both in the example set and the test sequences.



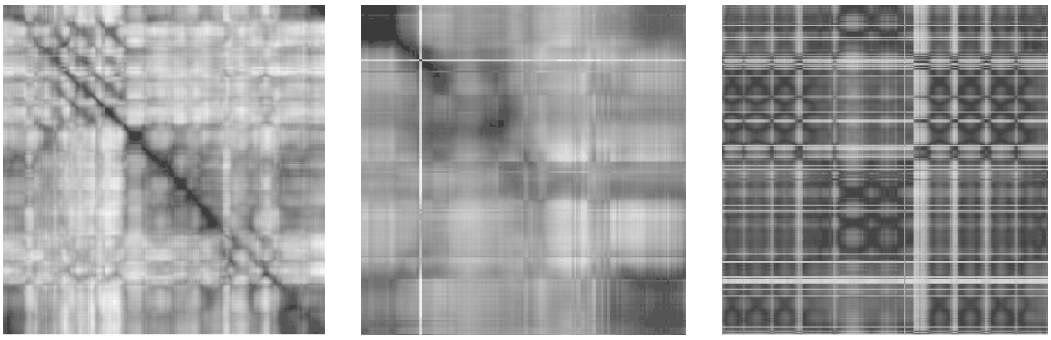
(a) Walking **(b)** Jog **(c)** Throw **(d)** Gesture **(e)** Box **(f)** Combo

Figure 3.4: Recovered poses of frames 100 of different actions, evaluated with a single camera using T1. (**a**,**d**) camera C1, subject 1 (**b**,**e**) camera C2, subject 2 (**c**,**f**) camera C3, subject 3. The erroneous estimation of the head tip (*headProximal*) marker in (**f**) is due to mistakes in the labelling of the training sequences of subject 3. The 3D pelvis location was set manually, as we only estimate relative joint locations.

The recovery accuracy for subject 3 is lower than for the other two subjects. There are two explanations for this. First, the location of the head tip (*headProximal*) marker is erroneous in some of the training sequences of subject 3 that are used in the example sets (see also Figure 3.4(f)). Given that many of the examples that are used in the recovery are from the same subject, this effect is stronger for the sequences of subject 3. Second, subject 3 wears dark clothes, which results in less prevalent edges between body and limbs. As such, forward-backward ambiguities are more likely to match well, in turn affecting the pose estimate. Also, differences in body dimensions play a role in the explanation of recovery accuracy. Since we use weighted interpolation of the joint positions, interpolation of similar poses of subjects with different body dimensions results in a different recovered pose. Naturally, differences are higher for joints that have a larger distance to the pelvis (*torsoDistal*) joint that was used for translation normalization.

We will discuss our results in more detail. If we look at different actions, we see large variations between sequences. In general, poses from the Walking and Jog sequences are recovered with the highest accuracy. This can, at least partly, be explained by the fact that these motions have been performed at least several times in the example sets. This ensures that more examples are available, thus increasing the probability of a close match. Moreover, each cycle resembles the others, in contrast to, for example, catching a ball where the ball appears at more or less random places. In Figure 3.5, we present the affinity matrices of walking cycles, catching and throwing a ball, and repetitions of waving and gesturing. For each matrix, a training and a test sequence from the same person are used. We immediately see the similarity between the two walking cycles by the dark line on the diagonal. This line is less apparent for two sequences of catching and throwing a ball. For the gesture sequences, it becomes clear that repetitions of the same gesture are very much alike. Also, the movement of

the limbs during walking and jogging was slower compared to the other actions, as noted by Bo and Sminchisescu [32]. This causes the edges to be more blurred, thus less distinctive. This might have had influence on the performance.



(a) Walking **(b)** Throw/Catch **(c)** Gestures

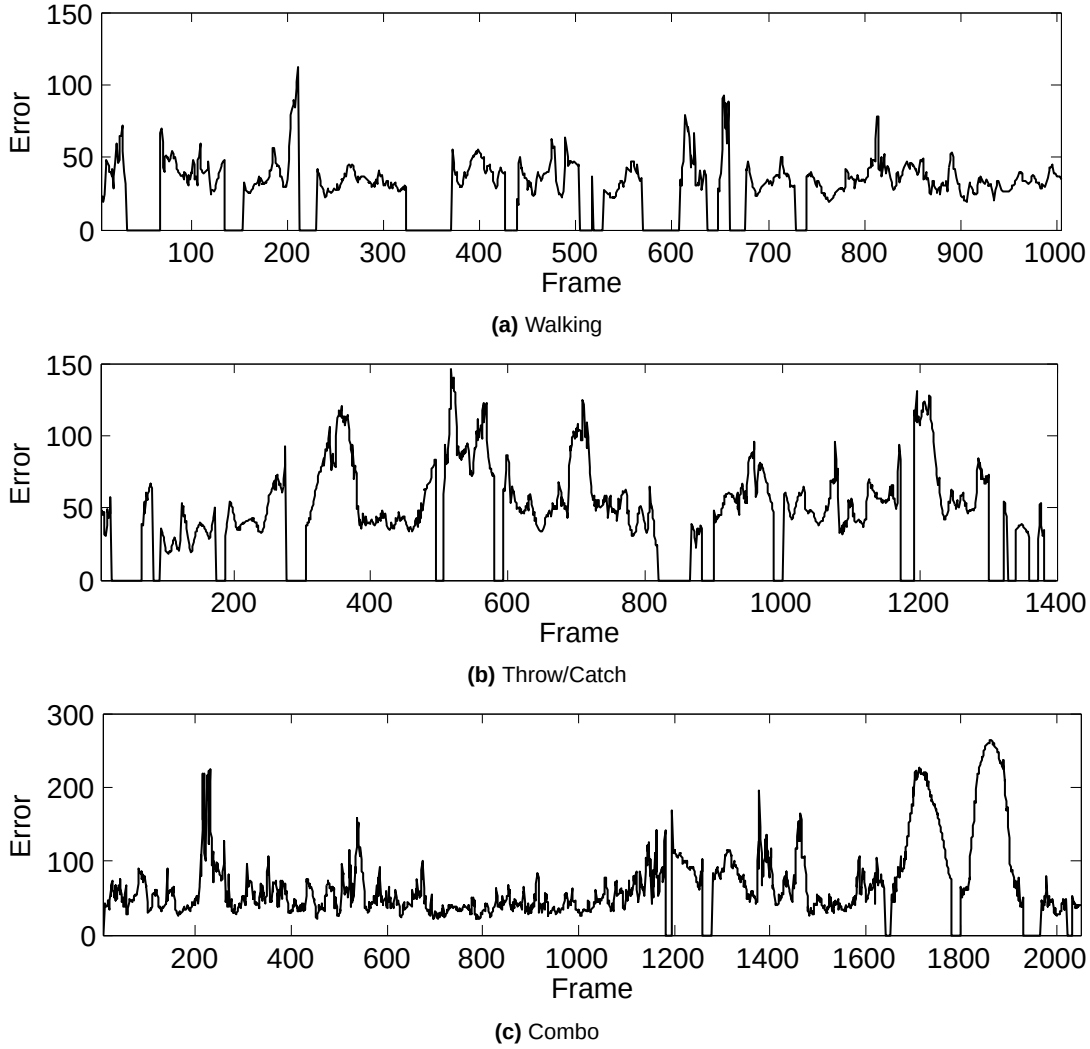
Figure 3.5: Affinity matrices for camera 1 of (**a**) Walking cycle, (**b**) Throw/Catch, and (**c**) Gestures sequence, performed by subject 1. Dark values correspond to small distances. Notice the similarity between the Walking cycles. In (**c**) corresponding areas are either repetitions of waving, or repetitions of the beckoning gesture. The white 'plus' in (**b**) is caused by incorrect segmentation due to the presence of the ball.

We should be able to see the same when looking at our test results. Figure 3.6(a–b) shows the plots of the mean errors over the sequences Walking and Throw/Catch, performed by subject 2 and obtained using all three cameras. The error plot is much more peaked for the Throw/Catch sequence. The peaks (e.g. around 350, 550 and 700) correspond to catching or throwing the ball. The lower errors around 450 and 650 correspond to waiting for the ball, which is in fact a standing pose. The slightly higher errors in the Walking sequence after 200, and around 650, correspond to the subject walking towards the camera. Here, depth ambiguities occur.

When we take another look at Table 3.2 and 3.3, we notice rather larger differences between subjects for the Gesture action. In this action, the subject waves and makes beckoning gestures. Subjects 1 and 3 perform these gestures with their right hand, in both the training and test sequences. This explains the low standard deviations. Subject 2 is more expressive in the performance, and is occasionally using both hands. Also, the body orientation in the test sequence is much more sideways than any of the sequences that has been used in the example set.

The Box actions show some of the highest errors, which is somewhat surprising. More careful analysis of the video data shows that, for subjects 2 and 3, there is quite some variation in the footwork. Subject 1 uses the same standing pose as a basis but is facing camera 1. From this view, it is difficult to estimate the depth of the arms. The view for cameras 2 and 3 are almost exactly from the side. This results in many estimates where the wrong arm is estimated to be stretched out.

The Combo action is a combination of walking, jogging, and some 'freestyle' moves (jumping on one leg and balancing on one foot). These moves are not present in the example set. For subject 2, the mean error over the whole sequence is shown in Figure 3.6(c). The peak around frame 250 is caused by incorrect background segmenta-



(a) Walking

(b) Throw/Catch

(c) Combo

Figure 3.6: Mean relative 3D error (in *mm*) plots for HumanEva-I (**a**) Walking, (**b**) Throw/Catch, and (**c**) Combo action, all performed by subject 2 and obtained using all three cameras. Instances that have a zero error contain invalid mocap.

tion. Frames 1160-1370 contain the jumping on one leg action, in frames 1660-1960 the subject balances on one foot with the arms stretched out. The difference in error is apparent, and gives us a clue about the performance of an example-based approach for unseen actions. We discuss this further in Section 3.2.4.2.

3.2.3.4 Results for HumanEva-II

The HumanEva-II set consists of two Combo sequences performed by subjects 2 and 4. HumanEva-II differs from HumanEva-I in the recording setup. Four color cameras are used, instead of the three color cameras and four grayscale cameras. The cameras are placed at different positions, and the elevation angles differ slightly. As mentioned before, this changed setup does not allow us to do any evaluation on the T3 example set.

Our evaluation was performed similar to that of the HumanEva-I test sequences. In Tables 3.4 and 3.5 the results are summarized for both subjects, for each camera separately. The three sets contain various movements within the sequence. Set 1 (frames 1-350) contains walking, set 2 (frames 1-700) contains walking and jogging, and set 3 (frames 1-1202 for subject 2, and 1-1258 for subject 4) contains the whole sequence. This includes walking, jogging, jumping on one leg and balancing on one foot.

	Set 1	Set 2	Set 3	Average
Camera 1	121.96 (72.51)	111.96 (59.83)	173.92 (111.22)	135.95 (81.19)
Camera 2	100.35 (34.28)	95.93 (34.54)	142.62 (77.93)	112.97 (48.92)
Camera 3	101.25 (44.72)	105.97 (60.93)	203.13 (144.72)	136.78 (83.46)
Camera 4	113.49 (48.19)	116.06 (54.85)	161.94 (89.45)	130.50 (64.16)
Average	109.26 (49.93)	107.48 (52.54)	170.40 (105.83)	126.71 (69.43)

Table 3.4: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-II test sequence Combo of subject 2, evaluated with a single camera. Results are broken down per set. Set 1 contains walking motions, set 2 contains both walking and jogging motions. Set 3 contains the entire sequence, including balancing motions.

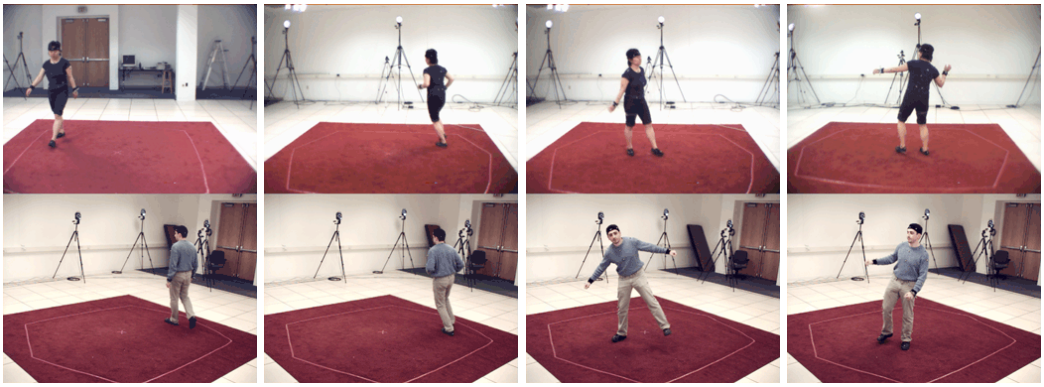
	Set 1	Set 2	Set 3	Average
Camera 1	129.92 (54.38)	138.89 (58.27)	166.31 (77.09)	145.04 (63.25)
Camera 2	120.73 (49.24)	115.93 (39.16)	147.50 (66.10)	128.05 (51.50)
Camera 3	183.61 (86.69)	146.75 (71.99)	200.31 (120.20)	176.89 (92.96)
Camera 4	146.60 (57.14)	151.44 (55.84)	203.77 (101.67)	167.27 (71.55)
Average	145.22 (61.86)	138.25 (56.32)	179.47 (91.27)	154.31 (69.82)

Table 3.5: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-II test sequence Combo of subject 4, evaluated with a single camera. Results are broken down per set. Set 1 contains walking motions, set 2 contains both walking and jogging motions. Set 3 contains the entire sequence, including balancing motions.

Our first remark concerns the analysis of the sequence performed by subject 4. We removed frames 298-337 from the results since these appeared to be erroneous. Specifically, we obtained mean errors per joint above 1,200 *mm*. Visual inspection of our pose estimates and the video revealed no peculiarities. Moreover, an average error per joint of 1,200 *mm* is very unlikely since we use relative distances.

Compared to the results that we obtained for the Combo sequences of HumanEva-I for T1, the errors for HumanEva-II are higher. Ideally, we would expect errors for set 1 that are comparable to those of the Walking sequences in HumanEva-I. Similarly, for set 2, we expected results that are close to those of Walking and Jog.

We observe that the example set does not contain any examples of subject 4. Differences in movement style and body dimensions between subjects can party explain the less accurate results. Also, although subject 2 appears in HumanEva-I, the clothing is different. This probably has an effect on the HOGs, and subsequently on the closest matches. In Section 3.2.4.2, we take a closer look at these issues. Yet, if we consider the subjects as unseen, the error is higher. We expect this to be the result of



(a) Frame 300 **(b)** Frame 600 **(c)** Frame 900 **(d)** Frame 1200

Figure 3.7: Single best estimate (top row) and original frame (bottom row) for the HumanEva-II Combo sequence performed by subject 2. Frame 300 shows forward-backward ambiguity in the walking action. Frames 900 and 1200 contain unseen movements.

the modified camera setup. The elevation of the camera is different, and there is a roll angle introduced for some of the cameras. We expect this to have quite a large impact, as HOG are not rotation invariant. Also, the background is modeled with a single Gaussian. This results in less accurate background segmentation, and this reflects on the HOG descriptors.

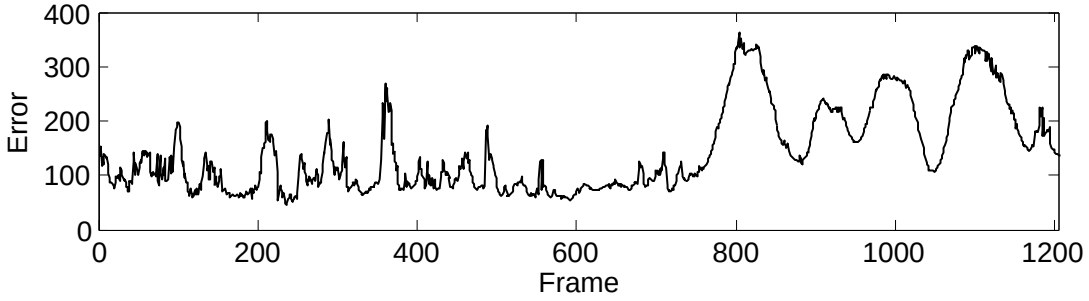
Figure 3.8 shows the mean error plots of the HumanEva-II sequences. The increased error for the unseen actions in set 3 (frame 750-end) is apparent. Other peaks (e.g. for subject 2 around frames 100, 220, 300 and 370) are due to forward-backward ambiguities. Figure 3.7 shows the original frames and the frames corresponding to the single best example.

3.2.3.5 Computational performance

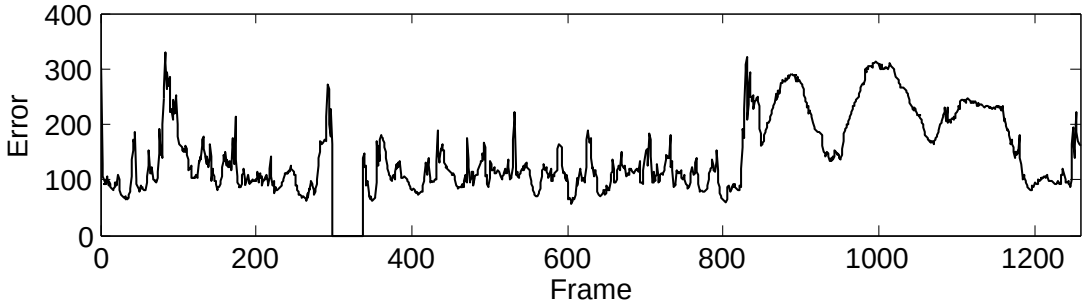
In this section, the computational performance for the different steps of our pose recovery approach are presented. We used un-optimized Matlab code, which was evaluated on a Pentium IV 2.8 GHz computer. For background subtraction, we used the code provided by Sigal and Black [322].

We used the walking test sequences as these present the largest variation in bounding box location due to distance of the subject to the camera. All images were 640×480 before applying background subtraction. A monocular setting with camera C1 and example set T1 was used, which contains 28,890 examples. Table 3.6 summarizes the computation times, which are averaged over subjects 1-3.

The most time-consuming step is the background subtraction of [322]. The background is described as a mixture of three Gaussians for each color channel. This results in many evaluations. In the current implementation, no look-up table is used, which causes the high computational cost. The step can be replaced by a more efficient implementation. Color conversion and shadow removal are only needed when the bounding box estimates are inaccurate. Using a different background subtraction algorithm can potentially solve this issue as well.



(a) Subject 2



(b) Subject 4

Figure 3.8: Mean relative 3D error (in *mm*) plots for the HumanEva-II Combo sequences performed by (**a**) subject 2, and (**b**) subject 4, both viewed with camera C1. Instances that have a zero error have been ignored (see text).

Process	Time (*ms*)
Background subtraction	1341.2
Color conversion	32.1
Shadow removal	57.0
HOG calculation	37.9
Finding best matches	285.5
Pose interpolation	0.7

Table 3.6: Computation times (in *ms*) for different steps in the pose recovery process. Times are averaged per frame over the Walking test sequences of subjects 1-3, viewed with camera C1 and evaluated using example set T1.

Given a segmented image, our un-optimized Matlab algorithm runs at 3 Hz on a standard PC. The HOG calculation itself is performed efficiently, which is partly due to the fact that only a fraction of the original image is used. The linear search for the closest matches is computationally demanding, while the interpolation time is negligible. In Section 3.2.4.4, we investigate the effect of a reduced number of examples on the performance. Instead of using a linear search, close examples can be found using hashing (e.g. [310]) or using a hierarchical matching procedure (e.g. [109; 403])

3.2.4 Additional experiments and results

In this section, we present additional experiments that give more insight in the capabilities in our proposed approach. To this end, we analyze the effect of HOG grid size and ROI setting in Section 3.2.4.1. Additional experiments where we remove different parts of the example set are presented in Section 3.2.4.2. We use these example sets further to analyze the performance of the HOG descriptor and the density of relevant examples in Section 3.2.4.3. Finally, we investigate the influence of the number of examples on the pose recovery accuracy in Section 3.2.4.4.

3.2.4.1 Results using different image representations

The ROI for the HOG-*F*-5×6 descriptor is determined by the minimum enclosing bounding box for the foreground pixels. The ROI is divided into a grid, and global normalization makes the descriptor effectively scale-invariant. In this process, the height and width are scaled independently, which can cause different poses to generate similar HOGs. See for example Figure 3.7(c) where, apart from the rotation, the height-width ratio is rather different. In this section, we introduce HOG-*R* descriptors, which are similar to the HOG-*F* descriptors but the ROI has a constant height-width ratio of 2.5. The figure is centered either horizontally or vertically in the ROI. The remaining space in the ROI is not part of the foreground mask.

Also, the number of columns and rows in the HOG-*F*-5×6 descriptor is an arbitrary choice, mainly motivated by the size of the head. In this section, we evaluate the performance for grid sizes 3×3, 4×4 and 5×6. The HOG-*R* descriptor sizes are 81, 144 and 270, respectively.

HOG descriptors are dense descriptors that take into account both the magnitude and the direction of all edges within a foreground mask. We expect that these edges help to partly encode depth in the pose. To investigate the effect of these edges, we introduce *histograms of oriented silhouette gradient* (HOSG) descriptors. These are essentially similar to the HOG descriptors, but only take into account silhouette boundary edges. Since these are binary and directed, we use 8 signed orientation bins, each of which covers 45°. The final descriptor sizes are 72, 128 and 240, for the HOSG-*R* descriptors with grid sizes of 3×3, 4×4 and 5×6, respectively. In the multi-view case, the length of the HOG and HOSG descriptors is tripled. Notice that we did not further investigate differences between HOG-*R* and HOG-*F* ROI settings.

The results for the different settings are presented in Tables 3.7 and 3.8, for the monocular and multi-view case, respectively. First thing to notice is the higher error of all settings, compared to those obtained with HOG-*F*-5×6 (see Tables 3.2 and 3.3). This is true for both the monocular, as the multi-view case. We expect that this effect is mainly caused by small variations in the extracted silhouette mask, especially in horizontal direction. A small change in this direction leads to a 2.5 times larger change in the vertical component. This is likely to cause pixels to be part of different grid cells.

Similar to earlier observations, we find that the results of the monocular case are less accurate than those of the multi-view setting. Especially the HOSG descriptors in the monocular case show relatively high errors. In the multi-view case, these errors

are more in line with those of the HOG descriptors. This finding can be explained by the fact the silhouettes cannot handle depth ambiguities well. Given only a single observation, this ambiguity cannot be resolved. This justifies the use of edges, but it should be noted that, for a given grid size, the HOG descriptor is of a slightly higher dimensionality than the HOSG descriptor.

When looking at the effect of grid size on the pose recovery accuracy, we observe a difference between HOG and HOSG. For HOSG descriptors, accuracy increases with grid size, both in the monocular as the multi-view case. However, for the HOG descriptors, the 4×4 setting appears to perform best. Differences in results between grid sizes are much smaller in the multi-view case. Probably, both the better results, as the more stable performance for different grid sizes, are due to the less frequent selection of ambiguous example. This also explains the lower standard deviation (not reported in the tables due to space constraints). These are approximately 25% lower for the multi-view case. Overall, these results show that reasonable results can be obtained by a relatively small descriptor. This is especially true when using edge information.

		HOG-*R*			HOSG-*R*		
Action	Subject	3×3	4×4	5×6	3×3	4×4	5×6
Walking	S1	59.64	49.14	76.31	91.63	76.75	71.13
Jog	S1	54.18	49.02	62.47	75.73	67.44	59.81
Throw/Catch	S1	N/A	N/A	N/A	N/A	N/A	N/A
Gestures	S1	25.07	23.57	28.08	27.45	25.96	25.37
Box	S1	89.94	76.10	91.46	87.58	82.06	82.47
Combo	S1	N/A	N/A	N/A	N/A	N/A	N/A
Average	S1	57.21	49.46	64.58	70.60	63.05	59.69
Walking	S2	44.87	42.13	58.91	87.41	68.68	57.98
Jog	S2	42.10	40.54	55.19	75.73	56.12	52.53
Throw/Catch	S2	70.10	70.96	91.05	95.03	84.18	80.93
Gestures	S2	82.69	87.19	91.54	98.82	93.85	88.09
Box	S2	114.86	118.06	134.60	139.34	135.28	121.63
Combo	S2	91.59	78.95	88.42	104.63	97.67	95.17
Average	S2	74.37	72.97	86.62	100.16	89.29	82.72
Walking	S3	65.54	61.16	74.56	82.26	74.84	70.11
Jog	S3	54.29	51.77	66.98	74.42	63.46	61.00
Throw/Catch	S3	112.44	117.11	131.13	112.52	111.29	116.72
Gestures	S3	59.34	64.16	76.61	59.88	61.47	59.07
Box	S3	114.53	101.05	114.83	118.73	113.56	112.99
Combo	S3	124.56	113.70	125.39	130.24	115.01	115.92
Average	S3	88.45	84.82	98.25	96.34	89.94	89.30
Average	All	75.36	71.54	85.47	91.34	82.98	79.43

Table 3.7: Mean relative 3D error in *mm* per joint for HumanEva-I test sequences, evaluated for different image representation settings, with a single camera (C1) using T1.

		HOG-*R*			HOSG-*R*		
Action	Subject	3×3	4×4	5×6	3×3	4×4	5×6
Walking	S1	40.72	38.43	44.22	56.00	47.00	43.74
Jog	S1	45.57	45.89	50.63	59.19	55.59	53.14
Throw/Catch	S1	N/A	N/A	N/A	N/A	N/A	N/A
Gestures	S1	22.74	24.36	24.42	23.81	24.69	24.06
Box	S1	76.78	87.91	83.27	86.32	87.80	94.64
Combo	S1	N/A	N/A	N/A	N/A	N/A	N/A
Average	S1	46.45	49.15	50.64	56.33	53.77	53.90
Walking	S2	40.93	41.29	42.32	45.02	43.64	43.17
Jog	S2	37.97	38.65	42.57	42.08	43.53	40.17
Throw/Catch	S2	57.32	58.28	61.00	70.21	65.39	66.71
Gestures	S2	81.87	74.93	78.88	97.01	93.53	92.61
Box	S2	94.80	95.33	94.32	103.37	103.82	105.21
Combo	S2	73.45	72.45	74.38	82.14	79.45	77.56
Average	S2	64.39	63.49	65.58	73.31	71.56	70.91
Walking	S3	54.59	57.11	59.30	60.51	58.00	59.67
Jog	S3	47.12	46.51	49.10	53.59	52.31	49.12
Throw/Catch	S3	92.76	93.92	106.90	90.42	88.47	85.39
Gestures	S3	56.83	54.80	59.63	59.13	56.18	61.34
Box	S3	103.61	94.28	106.73	116.07	120.35	116.50
Combo	S3	86.04	85.33	86.89	90.74	88.16	86.33
Average	S3	73.49	71.99	78.09	78.41	77.24	76.39
Average	All	63.32	63.09	66.54	70.98	69.24	68.71

Table 3.8: Mean relative 3D error in *mm* per joint for HumanEva-I test sequences, evaluated for different image representation settings and with all three cameras (C1–C3) using T3.

3.2.4.2 Results on validation set

In this section, we report on experiments where we removed parts of the example set. From the example set of HumanEva-I, we use the Walking sequence of subject 1 for testing. We choose the walking action as most work on human pose recovery has focussed on walking motion. As ground truth information is available for this sequence, we validate the sequences without using the online validation system. This allows us to analyze the accuracy performance of individual joint positions. Our experiments give insight in the degree of generalization over subjects and actions.

Since our test sequence is part of our example sets, we had to remove it. Our resulting example sets with the Walking trial of subject 1 removed are T1_T and T3_T, for the monocular and multi-camera cases, respectively. In addition, we created example sets where we removed all examples for the Walking action, T1_A and T3_A. This reduces the example set by 20%. Also, we removed all instances of subject 1, resulting in T1_S and T3_S, each containing roughly two thirds of the total number of samples (see Table 3.1). The results are presented in Table 3.9, with mean errors for the training part (frames 591-1203), the validation part (1-590) and over the entire sequence. This division into training and validation was suggested in [322]. We do

not distinguish between the two but report both numbers for comparison purposes.

Set	Train	Validation	Total
$T1_T$	80.14 (25.39)	74.40 (23.95)	77.15 (24.81)
$T3_T$	76.92 (28.50)	74.81 (23.86)	75.82 (26.20)
$T1_A$	94.72 (33.46)	92.69 (28.90)	93.66 (31.17)
$T3_A$	103.90 (46.23)	110.55 (44.96)	107.37 (45.68)
$T1_S$	84.18 (29.68)	76.86 (25.56)	80.36 (27.83)
$T3_S$	78.25 (29.02)	77.30 (24.12)	77.76 (26.57)

Table 3.9: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-I training sequence Walking, performed by subject 1. Errors are presented for different example sets.

Compared to the results in Tables 3.2 and 3.3, the errors obtained here are approximately 35-70 *mm* per joint higher. It should be noted that the number of examples is reduced and this is likely to have a negative effect on the pose recovery accuracy. We will further investigate this issue in Section 3.2.4.4. Our first observation is that even when only one trial is removed, the error is larger. Removal of this sequence results in a lower number of examples of the Walking action. Moreover, all examples of the same subject for the Walking action are left out. Apparently, our approach is more accurate when person-specific observations are used.

When all Walking examples are removed, the error increases 15-35 *mm*. Closer analysis of the examples used in the reconstruction shows that these are mainly from the Jog action. Walking and jogging show many resemblances, but in the Jog action the elbows are usually more bent, and the distance between the feet is kept small. The ankle and wrist joints indeed show a higher error.

A final observation can be made with respect to the removal of examples of subject 1. The results are only slightly less accurate than in the case where only the Walking action of subject 1 is removed. This leads us to conclude that only few examples from other actions are used in the estimation. Also, we have an estimate of the accuracy for pose recovery when the subject is not in the example set. Again, we must make a remark about the significantly reduced number of relevant samples, which we will further investigate in Section 3.2.4.4.

3.2.4.3 Results on validation set with pose matching

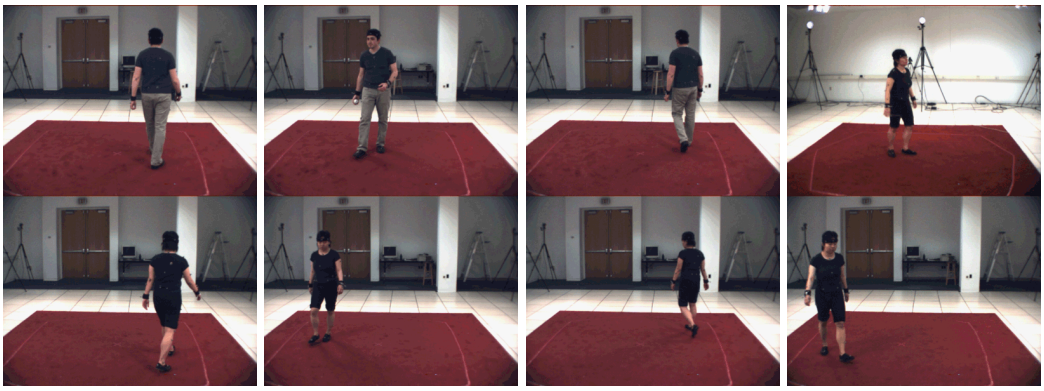
In our example-based pose recovery approach, we interpolated the poses of the closest HOG matches. Since the ground truth poses of the test set are known, we could perform the interpolation directly on the closest poses. Such an experiment gives insight in the performance of the HOG descriptor and the density of the pose space. We performed an experiment where we calculated the Euclidian distance between the pose of each test frame and the poses of all examples in the example set. We either selected the closest pose (1-NN) or used a weighted interpolation of the closest 25 poses (25-NN) as we would when matching HOG descriptors. We did not use any observation information and the notion of monocular and multi-view was therefore discarded. We used the poses of the three example datasets where we removed only the Walking trial of subject 1 ($T1_T$/$T3_T$), all walking sequences ($T1_A$/$T3_A$) or all

sequences of subject 1 ($T1_S$/$T3_S$). The results are presented in Table 3.10.

	$T1_T$/$T3_T$	$T1_A$/$T3_A$	$T1_S$/$T3_S$
1-NN	58.80 (7.68)	67.40 (8.45)	59.46 (7.93)
25-NN	55.29 (7.60)	61.09 (8.53)	58.01 (7.97)

Table 3.10: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-I training sequence Walking, performed by subject 1 when matching poses instead of HOGs. Errors are presented for different example sets.

Our first observation is that the use of interpolation results in a slight decrease of the error. Similar observations are reported in [274] when using image descriptors. Overall, the difference in error is relatively small. In Figure 3.9, several frames are shown together with the frame that corresponds to the closest matching pose from the example database.



(a) Frame 250 **(b)** Frame 500 **(c)** Frame 750 **(d)** Frame 1000

Figure 3.9: Frame corresponding to closest pose (top row) and original frame (bottom row) for the HumanEva-I Walking training sequence of subject 1 and evaluated using $T1_T$/$T3_T$. The closest poses originate from (**a**,**c**) the Walking sequence of subject 2, (**b**) the Throw/catch sequence of subject 2 and (**d**) the Throw/catch sequence of subject 1. The pose errors of the frames are 74.15, 59.56, 48.54 and 57.23 *mm*, respectively.

We compare the results in Table 3.10 with those in Table 3.9, obtained for HOG matching instead of pose matching and notice the lower error when using poses directly. If we would use an 1-NN approach using HOGs, the minimum error would be given by the reported 1-NN error using poses. Using HOGs, this error could only be obtained if the HOGs corresponding to the closest pose would also result in the lowest matching score. This is not always the case, as witnessed by the higher errors in Table 3.9. However, the errors are not that much higher, and the higher standard deviation could be caused by occasional mismatches. In Figure 3.10, the error graphs for the HOG matching and the 25-NN pose matching are shown. It becomes clear that, when using HOGs, there are indeed several peaks that increase both the average error and the standard deviation.

The errors that we obtain by selecting the closest examples of the validation set are higher than those obtained when matching HOGs on the Walking sequence of

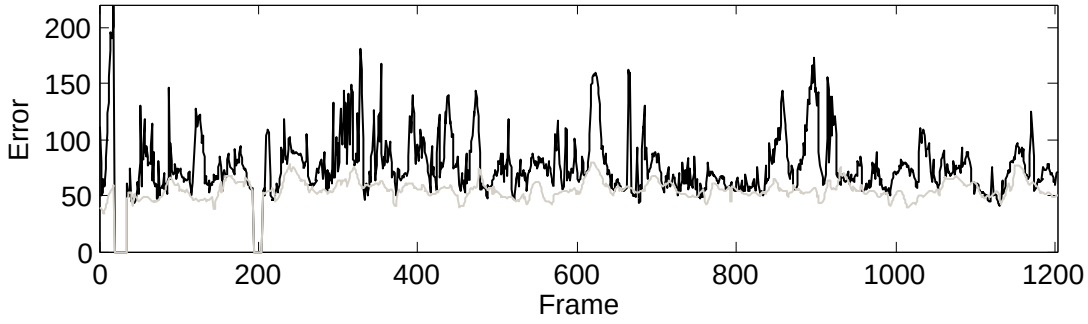


Figure 3.10: Mean relative 3D error (in *mm*) plots for HumanEva-I training sequence Walking, performed by subject 1 when matching poses (gray line) or HOGs (black line). In both cases, example set T1_T was used, and the estimates are obtained by weighted interpolation of the best 25 matches. Instances that have a zero error contain invalid mocap.

subject 1 in the test set (see Table 3.2). This might be explained by the presence of close examples in the T1 example set. These have been removed in example sets T1_T, T1_A and T1_S. But even when poses of the Walking sequences of other subjects are included in the example sets (as in T1_T and T1_S), the average pose error is still higher. We expect that this error can be explained by differences in body dimensions between subjects and partly by the differences in movement style. Given that the difference in body height between subjects 1 and 3 is larger than the difference between subjects 1 and 2, we expect that fewer of the closest pose matches are from subject 3. Indeed, in the 1-NN setting using T1_T, none of the closest matches are originated from sequences of subject 3.

3.2.4.4 Results using less examples

The nearest neighbor search has a computational complexity that is linear in the number of examples m. One straightforward way of reducing the computational cost would be to decrease m. In this section, we investigate the influence of the number of examples on the pose recovery accuracy. In contrast to the previous section, we remove examples from the full example set without taking into account the origin of the examples with respect to subject, action or camera view.

Specifically, we use k-medoids clustering [166], which is a variant of the k-means algorithm where, after each iteration, the cluster center is determined to be the closest example. After convergence, we retain those examples that are marked as cluster centers. We decided to cluster the poses, rather than the image descriptors. This decision is mainly practical, as the dimensionality of the poses is much lower. For large values of k, the calculation of the example-to-cluster distance matrix becomes practically infeasible due to the large space requirements. In our case, this happens for $k > 2000$. To overcome this issue, only in these cases, we select every l^{th} example from our original example set, where $l = \lfloor \frac{28{,}890}{k} \rfloor$. Given the high frame rate of the video (60 fps), the distance between subsequent frames is small. Therefore, we expect that the difference between the two sampling methods is marginal when l is small.

We evaluated the Walking sequences from the test set of HumanEva-I, subjects 1–3. The HOG-*F*-5×6 descriptors were used, calculated using only camera C1. Results are summarized in Table 3.11. It immediately becomes clear that pose recovery accuracy decreases with a decreasing number of examples. The effect is, however, not linear. With approximately 3,000 examples, a 10-fold reduction, the average error increases only by a factor 2. These results are in line with previous investigations of the influence of the number of examples on pose recovery [33; 357]. It should be noted that only approximately 30% of the examples originates from walking sequences. It is to be expected that the pose recovery accuracy can be increased by considering more examples, at the cost of higher computational demands.

Examples	Subject 1	Subject 2	Subject 3	Average
28,890	41.24 (16.84)	39.56 (26.75)	55.27 (21.70)	45.36 (21.76)
16,000	46.10 (18.56)	43.03 (26.46)	68.17 (21.38)	52.44 (22.13)
8,000	54.54 (21.04)	52.87 (26.20)	82.58 (24.69)	63.33 (23.98)
4,000	68.50 (23.09)	66.05 (28.76)	96.92 (31.98)	77.16 (27.94)
2,000	89.63 (28.99)	85.49 (38.55)	110.36 (41.42)	95.16 (36.32)
1,000	101.81 (34.73)	94.19 (39.02)	121.65 (42.98)	105.88 (38.91)
500	119.97 (31.15)	117.12 (36.07)	141.46 (41.26)	126.18 (36.16)
250	135.66 (31.66)	126.22 (31.62)	159.43 (43.45)	140.43 (35.58)
125	150.13 (31.20)	148.70 (35.41)	177.77 (36.21)	158.86 (34.27)

Table 3.11: Mean relative 3D error (and SD) in *mm* per joint for HumanEva-I Walking test sequences, evaluated with a single camera (C1) and with different example set sizes.

3.2.5 Discussion

In this section, we will briefly discuss our approach, and the results of the various experiments. By comparing these results to those reported for different approaches on the same dataset, we aim at revealing the strengths and weaknesses of the different image representations and recovery algorithms. Tables 3.12 and 3.13 summarize previous reports on HumanEva-I and HumanEva-II, respectively. We will first discuss how to interpret the results reported in the tables. Next, we will focus on discriminative approaches.

In Tables 3.12 and 3.13, the approaches are grouped by the error reporting manner. Errors can be either relative, or absolute. Relative errors, as we used in this section, are the relative to the pelvis (*torsoDistal*) joint. Global translations of the estimation of all joints, e.g. due to inaccurate person detection, will not affect this error. In contrast, the absolute error is indicative for both the person localization, and the recovered pose. In both cases, rotation of the person will affect the error. Reports can be either in 2D or 3D. In the former case, these are reported in pixels. The human figure is between 275 and 410 pixels in height, depending on the relative position to the camera. All 3D errors are reported in *mm*. In [80; 190], reported errors are normalized for rotation. It is unclear how much effect the normalization has on the reported accuracy. Several works report errors for joint angle estimates (e.g.

[28; 244; 410]) or present only quantitative results (e.g. [74; 316; 339]). As we cannot compare these result with ours, we did not include them in our comparison.

Care should be taken in directly comparing results. First, some authors have reported their results on the validation set. This is a valid approach, but it should be noted that the training and validation set are part of the same sequence. As a result, differences between training and validation sets might be smaller than between training and test sets. Moreover, the online validation system is not needed, as ground truth is available. This presents the risk of over-tuning parameters, especially when models are trained in a person-specific way. Some authors have used only part of the test sets. Especially for HumanEva-II, where the three parts are significantly different, this can have a big impact on the average score. For example, [60; 269; 293; 294; 303] evaluate only the part that contains walking movements. Second, some of the results are averages over multiple test subjects, whereas others are evaluated on a single subject (e.g. [135; 141; 190; 195]). As we have shown, accuracy of the recovered poses depends on the subject. Third, the training data that has been used in the various approaches differs to some extent. In some cases, training and testing has been performed specifically for a single person (e.g. [357]). Other authors have used very similar data that is not part of HumanEva, but has been used in [17]. A final issue is the difference between the joint positions that are used in the HumanEva dataset [322], and those that have been used in training the models [229; 272]. Differences in defined locations of the joint positions, relative to the body, result in systematic errors.

Tables 3.12 and 3.13 shows that comparable results are obtained for generative and discriminative approaches. Given the more flexible nature of generative approaches, one would expect better results. One possible explanation for the relatively good performance of discriminative approach is the controlled setting of the HumanEva set. Features such as silhouettes can be extracted relatively robustly. Moreover, the subjects that perform the test movements also appear in the training sets. For HumanEva-II, the situation is different and the recovery is clearly less accurate. This is true for both the generative and discriminative approaches, with the exception of Rosenhahn *et al.* [303] and Gall *et al.* [104]. They use a detailed 3D model that is obtained beforehand, and apply constraints that penalize self-intersections and intersections with the ground plane. It remains an open issue how detailed these models need to be, and how many constraints should be introduced to be able to achieve real-time performance. On the other hand, it also remains an unsolved issue how well discriminative approaches generalize to different settings.

For most of the approaches, the number of cameras does not seem to play a significant role. We already discussed that in our case, the multi-view setting performs slightly better, but this effect could be limited due to the small amount of available examples. In Husz *et al.* [141], the algorithm loses track when only a single view is used, which results in a significantly higher error for the monocular setting. It is difficult to explain why the monocular setting performs relatively well. We expect that the strength of edges as image representation is part of the explanation. Also, the use of a strong dynamical model (in combination with manual initialization) could be the reason that forward-backward ambiguities do not occur.

Overall, performance reported on HumanEva-II appears to be lower than results obtained on HumanEva-I sequences. In the previous section, we already discussed some of the possible explanations. First, the sequences in HumanEva-II contain balancing movements, which are present in the Combo sequences of HumanEva-I. These movements are usually recovered less accurately due to the lack of training data. Second, in HumanEva-II, one of the sequence is performed by a subject that has no training data available. A different movement style, and differences in visual appearance and body dimensions could explain the systematically higher errors for this subject. Another subject appears in both datasets, but wears different clothing which results in a different visual appearance. Third, the camera setup in HumanEva-II differs from that in HumanEva-I. The number of cameras and the viewpoint of each camera are different. Moreover, the sequences are much darker. In addition to a simpler standard background subtraction algorithm, this results in less precise foreground segmentations. Consequently, this affects the image representation.

3.2.5.1 Discussion of discriminative approaches

We will focus on the results obtained by discriminative approaches. First, various image descriptors have been used and evaluated. All of these rely on extracted silhouettes [33; 80; 135; 190; 293; 294] or edges [253; 272; 357]. In [135], attention is paid to accurate foreground segmentation, which is likely to increase the performance and generalization ability of the approach. However, silhouettes are prone to forward-backward ambiguities, and Howe [135] uses optical flow as an additional feature to achieve temporal consistency. When edges are used, depth ambiguities can be solved to some extent, but the representation becomes more person-specific due to strong edges present in the clothing. To this end, Okada and Soatto [253] discriminatively select those orientations within the HOG cells that are meaningful for a class of poses. When using their approach, no background subtraction is needed as background edges are effectively suppressed, similar to [3].

Another important factor is the use of dynamics. Elgammal and Lee [80] show that the application of a dynamical model improves the pose recovery performance. However, such an improvement comes at a cost of having to learn such a model. The authors learn a model in a projected 2D space, where one dimension models the viewing angle (or relative rotation of the person), while the other models the gait phase. Such a low-dimensional representation is suitable for cyclic actions such as walking and jogging, but it remains an open issue how to extend this work to more unconstrained actions such as throwing and catching a ball. Another disadvantage of using a dynamic model is the need for proper pose initialization. While this requirement is more prevalent in generative approaches to pose recovery, the risk is still present that errors are propagated through time due to incorrect initialization. Howe [135] avoids this problem by recovering a sequence of poses in batch. As such, initialization is not needed, but the approach cannot be used in real-time applications.

The discriminative approaches in Tables 3.12 and 3.13 are either example-based [135; 272] or regression-based [33; 80; 190; 293; 294; 357]. In an example-based approach, the pose space is implicitly described by the examples in the example set. Regression-based approaches describe the relation between image representation and

pose functionally. Due to the large number of pose variables and the high dimensionality of the image representation, such a mapping is usually described by a mixture of regression functions [26; 149]. Each of these regression functions, or experts, learns a mapping from part of the image space to the pose space. Gating functions are used to determine, based on the image representation, the probability that an expert provides the appropriate mapping. As noted in the previous section, and observed by Bo *et al*. [33] and Urtasun *et al*. [357], pose recovery accuracy increases with an increasing number of available examples. However, simultaneous learning of experts and gating functions becomes computationally inhibitive when the number of examples increases. This severely limits the application of a regression-based approach for pose domains that span several activities. Urtasun *et al*. [357] address the problem by learning local experts that cover only a very small portion of the image representation space. At test-time, for a given image representation, the nearest neighbors are determined, and the regression function is calculated using only the closest examples. This way, the mapping function can be learned effectively. The approach is related to the example-based approach, and is in fact a variant of locally weighted learning [14]. To allow for real-time recovery, the approach requires significant speed-up in the determination of the nearest neighbors. The authors propose to use hashing (e.g. [310]) to address this issue. Instead of using a partitioning of the image representation space, Bo *et al*. [33] propose algorithms that can efficiently learn conditional mixtures of experts. Traditionally, learning expert and gating functions is performed simultaneously, which requires a double-loop optimization approach. The authors introduce an approach which trains both models sequentially, which results in a decrease of both memory and computation requirements. Their algorithm thus can handle much larger numbers of examples.

3.3 Example-based pose recovery under partial occlusion

Given the common presence of partial occlusions due to other persons or the environment, there has been surprisingly little work that explicitly addresses this issue. Poppe and Poel [276] detect humans and recover their poses in single images. They use body-part templates and, for a match, vote over all joints in the human body. Such an approach can deal with severe occlusions, but is restricted to a limited class of motions (e.g. walking). Peursum *et al*. [268] use factored-state hierarchical HMM to model the motion of one given action. Occlusion of the feet can be detected using [267], and the likelihood function is adapted by ignoring the occluded area. The learned dynamical model will ensure stable tracking.

	Cams	Discr.	Image representation	Setting	Dyn.	Walking	Jog	Box	Abs.	2D/3D
Bo *et al.* [33]	1	Y	Histogram of SC	fBME-5	N	25.66	26.73	30.40	N	3D
Bo and Sminchisescu [32]	1	Y	HOG descriptor	TGPKNN	N	37.07	41.03	89.47	N	3D
Bo and Sminchisescu [32]	3	Y	HOG descriptor	TGPKNN	N	27.60	30.50	68.57	N	3D
Elgammal and Lee [80]	1	Y	Silhouette	Torus	Y	31.36			N	3D
Elgammal and Lee [80]	1	Y	Silhouette	Torus	N	34.58			N	3D
Lee and Elgammal [190]	1	Y	Silhouette	Torus inverse	N	76,56			N	3D
Okada and Soatto [253]	1	Y	HOG descriptor	6FSSVM + LRR	N	37.98			N	3D
Poppe [272] (this work)	1	Y	HOG descriptor	Nearest neighbor	N	45.36	43.92	94.48	N	3D
Poppe [272] (this work)	3	Y	HOG descriptor	Nearest neighbor	N	44.29	42.74	90.74	N	3D
Urtasun *et al.* [357]	1	Y	Hierarchical feature	GP-HF	N	32.70	31.20	38.50	N	3D
Li [195]	1	N	Chamfer distance		Y			187.50	Y	3D
Howe [135]	1	Y	Silhouette and flow	Temporal chain	Y	99.00			Y	3D
Mündermann *et al.* [229]	7	N	3D visual hull		Y	51.30		45.40	Y	3D
Ni *et al.* [237]	7	N	3D visual hull	Hybrid approach	Y	32.45	48.82	137.69	Y	3D
Peursum *et al.* [269]	3	N	Edge and silhouette	FSHHMM-PF	Y	92.89			Y	3D
Vondrak *et al.* [368]	3	N	Edge and silhouette	Physics	Y	93.40			Y	3D
Wu and Aghajan [395]	7	N	Edge and silhouette	NBP	N			93.18	Y	3D
Xu and Li [398]	7	N	Edge and silhouette	RBPF-PLS	Y	148.67			Y	3D
Zhang and Fan [407]	1	Y	Silhouette	MCMC-JD	Y	41.66			Y	3D
Howe [135]	1	Y	Silhouette	Temporal chain	Y	15.00			N	2D
Husz *et al.* [141]	1	N	Edge and silhouette	HPPF-AP	Y	38.10			N	2D
Urtasun *et al.* [357]	1	Y	Hierarchical feature	GP-HF	N	5.18	4.85	6.68	N	2D
Kuo *et al.* [182]	1	N	Edge, color, flow	Top 10	N	19.30			Y	2D
Poppe and Poel [276]	1	Y	Edge and color		N	27.48	30.35		Y	2D

Table 3.12: Comparison of results reported on HumanEva-I. For discriminative approaches (*discr.*), the image descriptor is mentioned, for generative methods, this is the image representation against which a model projection is matched. Dynamics (*dyn.*) indicates whether a dynamical model (possibly activity-specific) is employed. Absolute errors (*abs.*) include global translation errors. Relative errors are relative to the *torsoDistal* joint. Errors in 3D are in *mm*, for 2D in pixels. Direct comparison is hindered due to differences between evaluations (different subjects, validation set instead of test set, only part of the sequence). [80; 190] use a rotation-normalized error measure.

	Cams	Discr.	Image representation	Setting	Dyn.	Combo S2	Combo S4	Abs.	2D/3D
Cheng and Trivedi [49]	4	N	Voxel model	KC-GMM	Y	153.00	177.00	N	3D
Gall *et al.* [104]	4	N	Silhouette and color	ISA - 2 layers	Y	39.36	36.01	N	3d
Howe [135]	1	Y	Silhouette	Temporal chain	Y	126.75	166.00	N	3D
Husz *et al.* [141]	4	N	Edge and silhouette	HPPF-AP	Y	126.00	160.73	N	3D
Poppe [272] (this work)	1	Y	HOG descriptor	Nearest neighbor	N	126.71	154.31	N	3D
Cheng and Trivedi [49]	4	N	Voxel model	KC-GMM	Y	137.00	177.00	Y	3D
Darby *et al.* [60]	4	N	Edge and silhouette	APF	Y	145.00	147.00	Y	3D
Darby *et al.* [60]	4	N	Edge and silhouette	HMM+APF	Y	88.00	89.00	Y	3D
Gall *et al.* [104]	4	N	Silhouette and color	ISA - 2 layers	Y	37.53	32.01	Y	3d
Peursum *et al.* [269]	4	N	Edge and silhouette	FSHHMM-PF	Y	106.60	92.00	Y	3D
Rosenhahn *et al.* [303]	4	N	3D visual hull	Ground constraint	Y		33.80	Y	3D
Husz *et al.* [141]	4	N	Edge and silhouette	HPPF-AP	Y	21.76	23.14	N	2D
Rogez *et al.* [293]	1	Y	Silhouette		Y	17.75	15.62	Y	2D
Rogez *et al.* [294]	1	Y	HOG descriptor		N	13.58		Y	2D

Table 3.13: Comparison of results reported on HumanEva-II. For discriminative approaches (*discr.*), the image descriptor is mentioned, for generative methods, this is the image representation against which a model projection is matched. Dynamics (*dyn.*) indicates whether a dynamical model (possibly activity-specific) is employed. Absolute errors (*abs.*) include global translation errors. Relative errors are relative to the *torsoDistal* joint. Errors in 3D are in *mm*, for 2D in pixels. Direct comparison is hindered due to differences between evaluations (different subjects, only part of the sequence). [60; 293; 294; 303] evaluate only the walking part of the sequence.

For example-based approaches, occlusions are variations that are not explicitly modeled in the example set. To be able to handle these variations, there must be a way to take into account the missing (or ambiguous) information in the matching. To our best knowledge, only one paper takes into account occlusion in an example-based approach. Howe [134] uses boundary fragment matching to match partial shapes. Boundary fragments are small parameterized outlines of an extracted silhouette. Background and occlusion areas need to be labelled, so the matching algorithm knows which boundary fragments belong to the actual shape.

Instead, we use a slightly different approach as we focus on direct matching. This will guarantee a low computational cost. For the holistic HOG descriptors that we introduced in Section 3.2, local occlusions will affect the entire descriptor due to the global normalization. Moreover, matching is performed on the whole descriptor, which includes parts that contain occlusions. In Section 3.3.1, we adapt this matching to handle partial observations. Similar to Section 3.2, we assume that the foreground mask is available, and the ROI has been determined. In addition, we assume that occlusion can be predicted, similar to [134]. Specifically, this means that we know which parts of the ROI are occluded. We exploit the grid-based nature of the HOG descriptors to retrieve partial matches from the example set, where only cells without occlusion are used in the matching. This requires an adaptation of the global normalization to normalization per cell.

We explain in Section 3.3.1 which adaptations to our approach are required to allow for partial matches. Again, our approach is evaluated on the HumanEva dataset [322]. We perform a number of experiments with different simulated occlusion settings, which we compare to the results without occlusion. We describe these experiments in Section 3.3.2. A discussion of our approach and the results follows in Section 3.3.3.

3.3.1 Adaptations to the example-based pose recovery approach

In this section, we adapt the example-based pose recovery approach that we described and evaluated in Section 3.2. Our specific implementation differs on two aspects. First, in Section 3.3.1.1, we discuss the adapted normalization, which ensures that local changes do not affect the entire HOG descriptor. Second, in Section 3.3.1.2, we explain how we adapt our nearest neighbor search to handle partial observations. Other aspects of the approach remain unchanged. Specifically, the HOG-*F*-5×6 descriptors are calculated as in Section 3.2, with only edges in the foreground mask taken into account. We also weighted interpolation of the $n = 25$ closest examples to obtain the final pose.

3.3.1.1 Locally normalized HOG descriptors

In the calculation of the HOG descriptors, we only regard those pixels that are part of the human figure, and thus are part of the extracted foreground mask. In addition, we weight orientations according to their edge magnitude. This results in cells that are the weighted sums of edge orientations that correspond to pixels in the foreground.

HOG descriptors can be considered points in a 270-dimensional space, and can

be compared using e.g. Euclidean or Manhattan distance. To overcome differences in scale, clothing and lighting, we used global normalization to unit length in Section 3.2. However, occlusions cause some of the observation to be uninformative, or even misleading. Normalization of the entire descriptor dependents on all individual cells. In case of occlusion, some of the cells will have different edge responses. By normalizing descriptor d to unit length, these cells will affect all others. Therefore, we normalize each cell $h_{i,j}$ (i and j are a row and column index, respectively) individually to be of unit length. This has the advantage that we can still deal with variations in scale, as each cell is approximately equal in size. On the other hand, we discard the global character, and each cell contributes equally to the final descriptor. Specifically, individual cells that have a relatively low edge response, have a similar summed weight as cells that have a high response. Alternatively, we could have normalized each cell by its area size. This would make the descriptor invariant to scale, but would not take into account variations due to different lighting settings and clothing.

3.3.1.2 Matching of partial HOG descriptors

We introduce weights $\phi_{i,j}$ for each cell. These weights scale the feature space, and therefore affect the distance functions that we define. While these weights can take any value (e.g. in the $[0, 1]$ range), we use here binary values. That is, $\phi_{i,j} = 0$ in the case of occlusion within the cell, and $\phi_{i,j} = 1$, otherwise. This weighting effectively determines a lower-dimensional subspace, in which we can perform matching. The distance between descriptor d with cell-normalized histograms $\hat{h}_{i,j}$ and descriptor d' with cell-normalized histograms $\hat{h'}_{i,j}$ is calculated as:

$$D(d, d') = \frac{\sum_{i=1}^{5} \sum_{j=1}^{6} \phi_{i,j} \delta(\hat{h}_{i,j}, \hat{h}'_{i,j})}{\sum_{i=1}^{5} \sum_{j=1}^{6} \phi_{i,j}} \tag{3.1}$$

Here, n_r and n_c are the number of rows and columns, respectively. For HOG-*F*-5×6, $n_r = 5$ and $n_c = 6$. δ is the distance measure between two histograms, for which we again use Manhattan distance. Note that, since we predict which cells are partly occluded, we could ignore these cells and normalize the descriptor to unit length. This approach would maintain the advantages of a global normalization but has the drawback that normalization of all n examples in the database needs to be performed at run time. This will severely affect the computational performance.

3.3.2 Experiment results

To our best knowledge, there is no dataset that contains both ground truth motion capture data, and video data with occlusions. Therefore, we use the HumanEva dataset, and simulate different types of occlusion. Specifically, we use the HOG-F-5×6 descriptors and the monocular example set T1, as described in Section 3.2.3.2. The test sets are presented in Section 3.3.2.1. Results are presented in Section 3.3.2.2 and discussed in Section 3.3.3.

3.3.2.1 Test sets

Similar to our experiments without occlusion, we present results for all test trials of subjects 1, 2 and 3 in HumanEva-I. Evaluation of the Combo and Throw/Catch test sequences of subject 1 failed repeatedly. Consequently, we can not report our results on these trials. Note that part of the Combo sequences contain movements that are not in the example set, which our algorithm cannot handle.

We do not have access to data that contains video of subjects under occlusion, synchronized with corresponding pose information. Therefore, we simulate occlusion on the HumanEva-I dataset. We define six different occlusion settings. Each of these settings is a fixed combination of weights $\phi_{i,j}$. Effectively, we simulate occlusion on a fixed set of cells, not regarding the location of the subject in the image. As such, we can test the influence of occlusion on a large number of poses. The six settings that we define are *v_left*, *v_center*, *v_right*, *h_top*, *h_center* and *h_bottom*, see also Figure 3.11. In the vertical and horizontal conditions, 20% and 33% of the image observation is occluded, respectively.

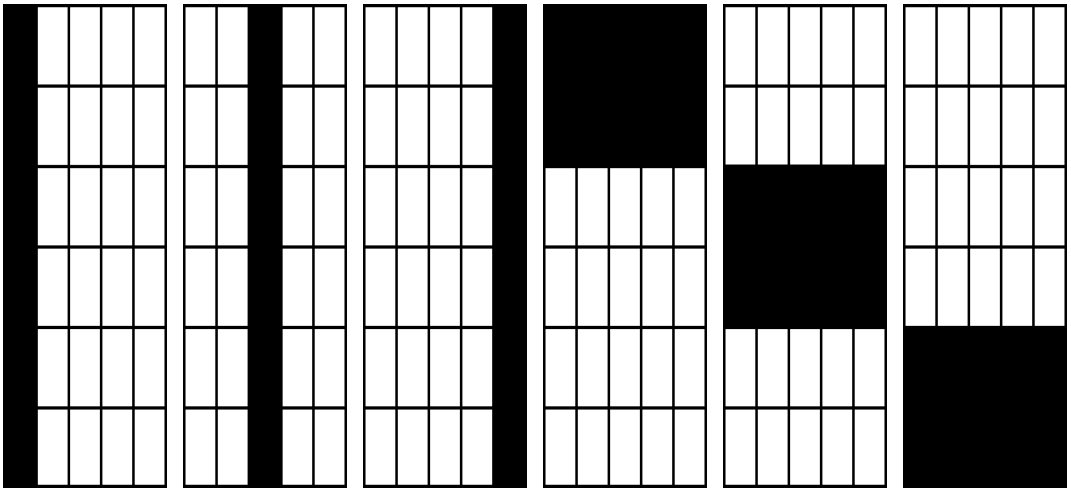


Figure 3.11: Occlusion settings (left to right) *v_left*, *v_center*, *v_right*, *h_top*, *h_center* and *h_bottom*

In addition to the fixed occlusion settings, we also created a *pole* sequence. This is the walking sequence of subject 1, but with a fixed occluded area in the image, not relative to the ROI, see also Figure 3.12(a). The area is a vertical pole with a width of 40 pixels. For comparison, the subject in the image is on average approximately 115 pixels in width. Due to the scale and pose, this can vary between 70 and 170 pixels. The *pole* sequence is a more realistic scenario and can show how the estimation error changes as the subject moves through an occlusion.

3.3.2.2 Results

Table 3.14 summarizes the average 3D errors over all joints, relative to the pelvis. The last column shows the results without occlusion. These numbers differ from those reported in Table 3.2, due to the different normalization. On average, the

cell-based normalization results in slightly higher average error compared to global normalization (68.92 and 65.63 *mm*, respectively). For each occlusion condition, the table shows the increase in error over all evaluated actions and subjects. For the vertical conditions, each of which occludes 20% of the image, the average error is approximately 7% higher. The horizontal settings result in a 9% increase, while occlusion covers one third of the observation. There are, however, large differences in performance for different trial. We will discuss these in the next section.

For the *pole* sequence, we obtained a 3D error of 43.54 *mm* averaged over all joints, and relative to the pelvis joint. We also discuss these results in the next section.

Action	Vertical			Horizontal			None
	Left	Center	Right	Top	Center	Bottom	
S1 Walking	39.80	43.28	42.10	47.52	43.83	45.83	39.03
S1 Jog	58.24	57.16	54.09	59.34	52.27	63.42	48.75
S1 Throw/Catch	N/A	N/A	N/A	N/A	N/A	N/A	N/A
S1 Gestures	30.72	28.93	30.75	30.58	30.44	31.79	30.11
S1 Box	84.40	93.13	84.80	97.81	90.30	79.41	81.38
S1 Combo	N/A	N/A	N/A	N/A	N/A	N/A	N/A
S2 Walking	37.60	40.18	39.22	42.23	39.53	41.24	37.27
S2 Jog	43.11	41.66	43.34	49.00	45.16	50.79	41.37
S2 Throw/Catch	73.53	72.13	71.34	76.81	74.06	76.33	72.18
S2 Gestures	95.30	95.65	86.14	84.90	85.25	93.16	82.81
S2 Box	109.17	120.03	107.43	107.14	112.11	115.45	105.08
S2 Combo	71.20	72.72	72.67	78.30	74.14	76.05	68.73
S3 Walking	58.38	64.73	64.23	61.31	61.22	65.79	58.86
S3 Jog	52.57	54.55	50.77	53.26	52.55	53.63	47.83
S3 Throw/Catch	120.06	112.22	94.38	110.81	103.15	122.22	101.40
S3 Gestures	80.09	76.85	97.63	72.49	91.20	114.00	82.63
S3 Box	105.64	110.48	98.69	110.15	105.65	106.99	98.44
S3 Combo	117.75	116.97	111.11	119.18	112.84	116.02	106.85
Average	73.60	75.04	71.79	74.43	73.36	78.26	68.92
Increase (%)	6.79	8.88	4.16	7.99	6.44	13.56	0.00
Occluded (%)	20.00	20.00	20.00	33.33	33.33	33.33	0.00

Table 3.14: Mean relative 3D error in *mm* per joint for HumanEva-I test sequences, evaluated with a single camera (C1) using T1. For each fixed occlusion setting, the increase over the non-occluded observations and the amount of occlusion are given.

3.3.3 Discussion

Several sequences show slightly lower errors when occlusion is added. This is most likely caused by bad foreground segmentation. In the occluded conditions, these distracting regions are ignored.

In general, we observe differences in accuracy between occlusion settings. The horizontal settings have a slightly higher error (9%), compared to the vertical occlusion settings (7%). This effect can be partly explained by the higher percentage of

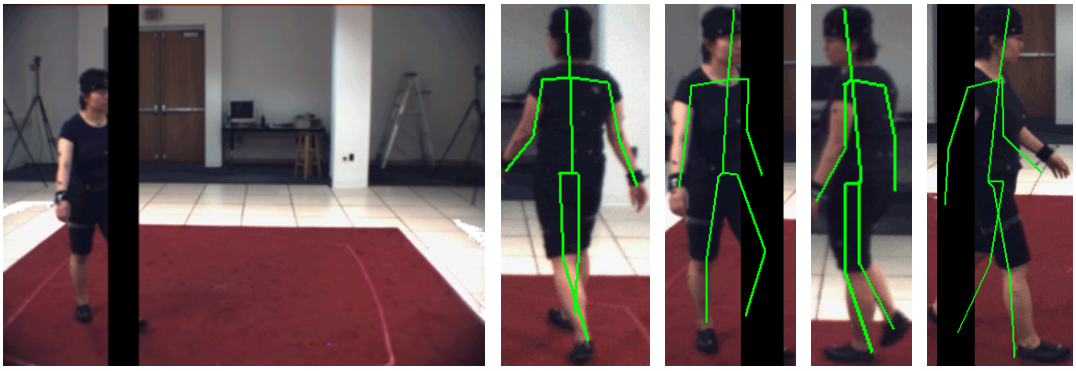


Figure 3.12: (**a**) Full frame of the *pole* sequence, with (**b**-**e**) recovered poses of frames 250, 500, 750 and 1000. In frames 500 and 1000, the pole is shown, but a larger region is not taken into account (60 and 40%, respectively). The 3D pelvis location was set manually, as we only estimate relative joint locations.

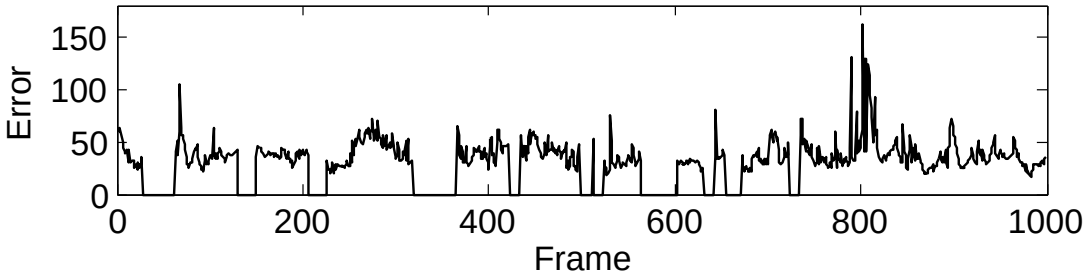
occlusion (33%, compared to 20%). However, there are differences between settings. Most of these differences are caused by several individual trials. We will discuss these sequences that largely contribute to the increased error.

A significant part of the increase in recovery error is due to the boxing action. Especially for subject 3, almost all occlusion settings result in significantly higher errors. Analysis of these results reveal that many examples from the gesture action are selected. For the gesture action, a similar trend is visible, especially in the *h_bottom* condition. We expect that this is mainly due to the fact that subject 3 wears dark clothes, which results in relatively few edge responses between arms and body. Therefore, examples from beckoning gestures and box punches are selected interchangeably. The same effect is visible for subject 2, but to a much lesser extent.

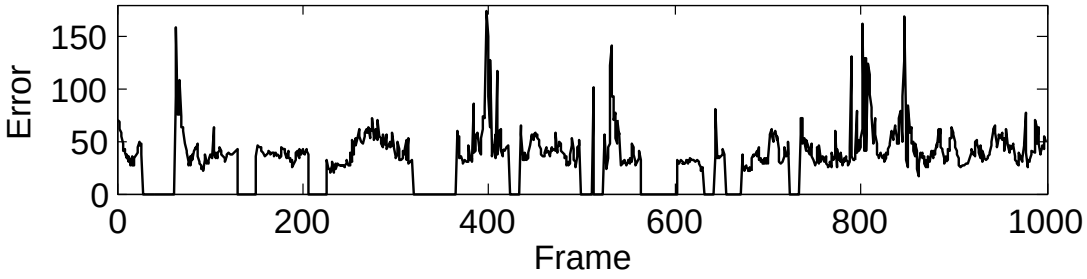
For the jog action, recovery accuracy is significantly lower in the *h_bottom* occlusion setting. As the hands do not move a lot while jogging, the legs are most information in this case as well.

The throwing and catching trials of subject 2 have a relatively low increase in error, whereas the *h_bottom* condition for subject 3 shows much higher errors. Similar to earlier observations, the relatively low number of edge responses for subject 3 is probably the reason that the missing edges for the feet result in decreased accuracy. Also, the *v_left* occlusion setting shows higher error values, which can be explained since the subjects both throws and catches the ball right to him. In the image, this corresponds to the left side of the observation.

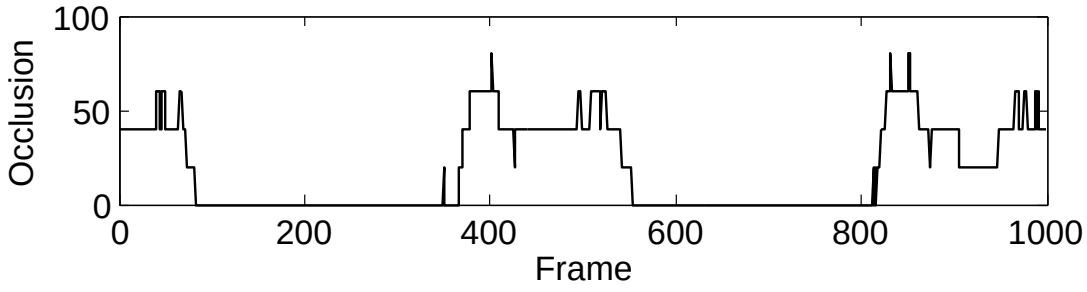
The error plots for the *pole* sequence are shown in Figure 3.13. The amount of occlusion for each frame is in the range 0-80%. On average, 19.06% of the observation is occluded. The graphs clearly show that the error increases under occlusion. The average increase in error is 11.56%. It should be noted that this is mainly due to those frames where more than half of the observation is occluded. Here, the average error is 57.00 *mm*, compared to 39.17 *mm* for the non-occluded sequence. When occlusion is in the range 20-40%, the mean error is 43.63 *mm*, compared to 37.49 *mm*



(a) No occlusion



(b) Occlusion simulated (*pole* sequence)



(c) Percentage of occlusion

Figure 3.13: Mean relative 3D error (in *mm*) plots for HumanEva-I Walking, performed by subject 1 and viewed with a single camera (C1), (**a**) without occlusion, (**b**) for *pole* condition and (**c**) percentage of occlusion for *pole* condition. Instances that have a zero error contain invalid mocap.

for the original. Differences between the upper two graphs for those frames where no occlusion is present are due differences in normalization. The pose estimates for several frames are shown in Figure 3.12(b-e).

In summary, we have shown that, when the occlusion areas are known, we can still recover poses. We obtained an average increase in recovery error of 6% when using vertical occlusions that cover 20% of the observation. Horizontal occlusions that cover one third of the ROI result in an increase of approximately 11%. Apart from the novelty of using a direct matching approach for poses under occlusion, we have shown that moderate occlusion sizes only increase the pose recovery accuracy slightly. The drawback of our approach is that we need to predict the occlusion areas, and require a relative good fit of the ROI.

Part II

Human action recognition

4

Human action recognition: an overview

4.1 Introduction

We consider the task of labeling videos containing human motion with action classes. The interest in the topic is motivated by the promise of many applications, both offline and online. Automatic annotation of video enables more efficient searching, for example finding tackles in soccer game recordings, handshakes in news footage or typical dance moves in music videos. Online processing allows for automatic surveillance, for example in parking lots or shopping malls, but also in smart homes for the elderly, to support aging in place. Also, interactive applications, for example in human-computer interaction or games, could benefit from many of the advances in automatic human action recognition.

A large body of research has been carried out, mainly in the last couple of years. Part of this work has already been adopted in commercial applications, such as video-based game controllers and sport video analysis. Many of the recent trends appear promising for increasingly challenging domains, and are the focus of this survey.

In this section, we first discuss related surveys, and present the scope of this overview. Also, we outline the main characteristics and challenges of the field, as these motivate the various approaches that are reported in literature. Finally, we briefly describe the most common datasets.

In its simplest form, vision-based human action recognition can be regarded as a combination of feature extraction, and subsequent classification of these image representations. We discuss these two tasks in Sections 4.2 and 4.3, respectively. Many works will be described and analyzed in more detail. However, we do not intend to give complete coverage of all works in the area. Finally, we discuss limitations of the state of the art and outline future directions to address these in Section 4.4.

4.1.1 Scope of this overview

The area of human motion recognition is closely related to other lines of research that analyze human motion. Moreover, the recognition of action can be performed at various levels of abstraction. Different taxonomies have been proposed and here

we adopt the hierarchy used by Moeslund *et al*. [222]: action primitive, action and activity. An action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the action that is being performed. For example, 'putting left leg forward' is an action primitive, whereas 'running' is an action. 'Jumping hurdles' is an activity that contains starting, jumping and running actions.

We focus on actions and do not explicitly consider context such as the environment (e.g. [290]), interactions between persons (e.g. [263; 304]) or objects (e.g. [120; 224]). These approaches fall outside the scope of this overview. Moreover, we consider only full-body movements. This excludes the work on gesture recognition, for which the reader is referred to surveys by Mitra and Acharya [220], and Erol *et al*. [83].

One important consequence of our focus is that we group many different performances of movement under the same action category. This is an arbitrary process as there is often significant intra-class variation. We discuss this issue in more detail in Section 4.1.3. Here, we explicitly discuss the differences between human action recognition and human gait recognition, which is an established field of research. While gait recognition focusses on identifying different styles of walking movement, the aim in human action recognition is the opposite: to generalize over these variations. Recently, there have been several approaches that aim at simultaneous recognition of both action, and style (e.g. [57; 78; 370]). In this overview, we will discuss mainly those approaches that can deal with a variety of actions with different spatial and temporal characteristics.

4.1.2 Surveys and taxonomies

There are several existing surveys within the area of vision-based human motion analysis and recognition. Recent overviews by Forsyth *et al*. [97] and Poppe [273] (see also Chapter 2) focus on the recovery of human poses and motion from image sequences. This can be regarded as a regression problem, whereas human action recognition is a classification problem. Nevertheless, the two topics share many similarities, especially at the level of image representation. Also related is the work on human or pedestrian detection, where the task is to localize persons within the image. Surveys can be found in [82; 106].

Broader surveys that cover the above mentioned topics, including human action recognition, appear in [5; 34; 108; 180; 222; 351; 373]. Bobick [34] uses a taxonomy of movement recognition, activity recognition and action recognition. These three classes correspond roughly with low-level, mid-level and high-level vision tasks. It should be noted that we use a different definition of action and activity. Aggarwal and Cai [5], and later Wang *et al*. [373], discuss body structure analysis, tracking and recognition. Recognition is further divided into template matching and state-space approaches. Gavrila [108] uses a taxonomy of 2D approaches, 3D approaches and recognition. Moeslund *et al*. [222] use a functional taxonomy with subsequent phases: initialization, tracking, pose estimation and recognition. Within the recognition task, scene interpretation, holistic approaches, body-part approaches and action

primitives are discussed. A recent survey by Turaga *et al*. [351] focuses on the higher-level recognition of human activity. Krüger *et al*. [180] additionally discuss intention recognition and imitation learning.

Overall, these surveys cover a broad range of work from human localization to behavior interpretation. This consequently makes it harder to address typical issues of a single domain. In contrast, we focus on vision-based human action recognition only and address characteristics and challenges explicitly (see Section 4.1.3). Instead of using a functional taxonomy, we discuss image representation and action classification separately as these are the two parts that are present in every action recognition approach. Such a taxonomy allows to focus on typical issues and characteristics. Due to the large variation in datasets and evaluation practice, we discuss action recognition approaches conceptually, without presenting detailed results. We focus on recent work, which has not been discussed in previous surveys. In addition, we present a discussion that focusses on promising work and points out future directions.

4.1.3 Challenges of the domain

In human action recognition, the task is to analyze video and to issue a corresponding class label. We identify challenges due to differences in performance, and due to differences in the recording. In this section, we discuss these in more detail.

4.1.3.1 Intra- and inter-class variations

For many actions, there are large variations in performance. For example, walking movements can differ in speed and stride length. Also, there are anthropometric differences between individuals. In fact, personal differences in gait have motivated its use as a biometric cue. Similar observations can be made for other actions, especially for non-cyclic actions or actions that are adapted to the environment (e.g. avoiding obstacles while walking, or pointing towards a certain location). For multiple classes, distinguishing becomes more challenging when the intra-class variation of each class is high. For example, slow running resembles jogging. A good human action recognition approach should be able to generalize over variations within one class, while at the same time to distinguish between actions of different classes.

4.1.3.2 Environment and recording settings

Even when actions are performed in the same manner, differences in the recording setup and environment result in differences in the captured movement. Since we focus on vision-based human action recognition, we address these differences explicitly. The environment in which the action performance takes place, is an important source of variation in the recording. When this environment is cluttered or dynamic, it might prove harder to localize the person in the video. Moreover, the environment or recording setup might be such that parts of the person might be occluded in the recording. This introduces a source of uncertainty.

Also, the fact that a single camera is only able to capture a projection introduces a source of variation. The same action, observed from different viewpoints, can lead to very different image observations. Often, a known camera viewpoint is assumed,

but this restricts the use to static cameras. When multiple cameras are used, viewpoint problems and issues with occlusion can be alleviated, especially when observations from multiple views can be combined into a consistent representation. Dynamic backgrounds further increase the complexity of localizing the person in the image and robustly observing the motion. When using a moving camera, these issues become even harder.

Different persons can appear differently due to differences in anthropometry, but also due to clothing, skin color and facial appearance. Lighting conditions can further influence the appearance. A robust approach should be able to generalize over these factors or employ an initialization phase.

4.1.3.3 Spatial and temporal variations

Since human motion is performed by a person, a common approach is to localize the person in the image or video first. There may be variations in the localization, and human action recognition algorithms should be able to cope with these.

There can also be variation in the detection in the temporal domain. Often, actions are assumed to be segmented in time before the actual action classification takes place. Such an assumption moves the burden of the segmentation from the recognition task, but requires a separate segmentation process to have been employed previously. This might not always be realistic.

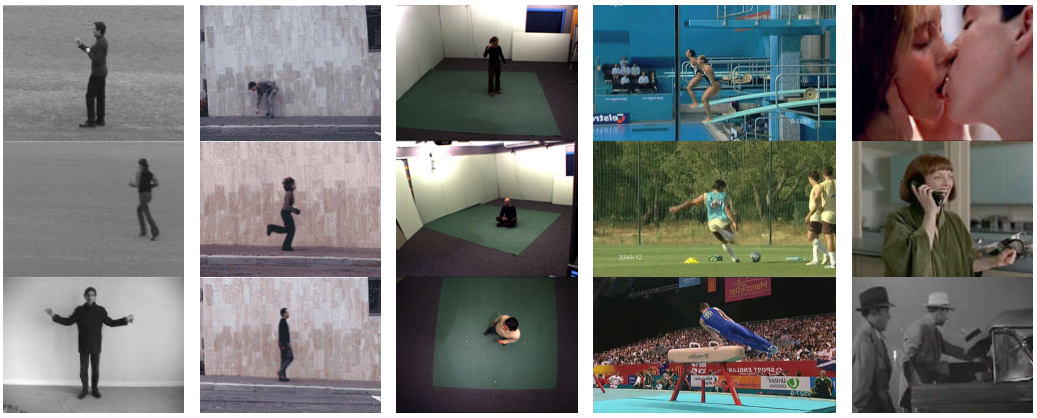
Also, there is substantial variation in the rate of performance of an action. We already discussed inter-personal variations, but the rate at which the action is recorded also has an important effect on the temporal extent of an action, especially when motion features are used. A robust human action recognition algorithm should be invariant to these different rates of execution.

4.1.3.4 Evaluation criteria

Within the domain, much of the evaluation efforts are focussed on publicly available datasets (see also Section 4.1.4). While this provides a sound mechanism for comparison, there is the risk of tuning the algorithms to the datasets. In particular, the (un)availability of several of the above mentioned variations strongly guides design decisions. Therefore, public databases with sufficient variation for these challenges are necessary. We will discuss this issue in more detail in Section 4.4. Related is the reliability of the data labeling. Most existing data uses actors that perform predefined actions. This readily provides the data labeling. However, performance of an action might be perceived differently by different people. A small-scale experiment was performed in [264], and showed significant disagreement between human labeling and the assumed ground-truth on a common dataset. When no ground truth is available, an unsupervised approach needs to be pursued. While such an approach will discover classes of similar movement, there is no guarantee that these classes are semantically meaningful.

4.1.4 Common datasets

The use of common, publicly available datasets allows for the comparison of different approaches and gives more insight into the (in)abilities of respective methods. A number of datasets have been recorded and made available to the general audience. We discuss these sets and their characteristics below.



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

Figure 4.1: Example frames of (**a**) KTH dataset, (**b**) Weizmann dataset, (**c**) Inria XMAS dataset, (**d**) UCF sports action dataset and (**e**) HOHA dataset.

4.1.4.1 KTH human motion dataset

The KTH human motion dataset (Figure 4.1(a), [307]) contains 6 actions (walking, jogging, running, boxing, hand waving and hand clapping), performed by 25 different actors. Four different scenarios are used: outdoors, outdoors with zooming, outdoors with different clothing and indoors. For each combination of person and scenario, each action is performed four times sequentially. Some sequences are missing, which brings the total number of performances to 2391.

The sequences are 160 $\times$ 120 pixels, with the person approximately 50–80 pixels in height. The recordings are at 25 frames per second, and each repetition lasts on average 4 seconds. There is considerable variation in the performance and duration, and somewhat in the viewpoint. The walking, jogging and running actions are performed either from left to right, or in the opposite direction. The backgrounds are relatively static, but hard shadows are usually present. In addition, some of the recordings are overexposed. Apart from the zooming scenario, there is only slight camera movement.

4.1.4.2 Weizmann human action dataset

The human action dataset (Figure 4.1(b), [30]) recorded at the Weizmann institute contains 9 actions (walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place and jumping jack), each performed by 10 persons. A skip action was added later. Each sequence is 180 $\times$ 144 pixels with a subject height

of approximately 60 pixels. The duration of a sequence is 2–3 seconds on average and recorded at 25 frames per second. The backgrounds are static and foreground silhouettes are included in the dataset. There is some variation in the performance of the actions between persons. The viewpoint is static but some of the actions are performed either from left to right, or in the opposite direction.

In addition to this dataset, two separate sets of sequences were recorded for robustness evaluation. One of these sets shows walking movement viewed from different angles. The second set shows fronto-parallel walking actions with slight variations (carrying objects, different clothing, different styles).

4.1.4.3 INRIA XMAS multi-view dataset

Weinland *et al*. [388] introduced the IXMAS dataset (Figure 4.1(c)) that contains actions captured from five viewpoints. A total of 11 subjects perform 14 actions (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw over head and throw from bottom up). The actions are performed in succession, usually with a short period between two actions. Each subject performed three trials. There is considerable variation in action performance. Notably, the actions are performed in an arbitrary direction with regard to the camera setup.

The sequences are recorded at 390 $\times$ 291 pixels, 23 frames per second. The subject is approximately 120–140 pixels tall in the images. The camera views are fixed, with a static background and illumination settings. This allows for robust background subtraction, and silhouettes are part of the dataset. Also, synchronization of the cameras allowed for the construction of volumetric voxel (64 $\times$ 64 $\times$ 64) representations, which are also included.

4.1.4.4 UCF sports action dataset

The UCF sports action dataset (Figure 4.1(d), [291]) contains 150 sequences of sport motions (diving, golf swinging, kicking, weight-lifting, horseback riding, running, skating, swinging a baseball bat and walking). Bounding boxes of the human figure are provided with the dataset. All sequences have a resolution of 720×480 or 720×404 and are collected from broadcast television channels. Each sequence is recorded at 25 frames per second, and is on average 3 seconds long. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and backgrounds. Some of the sequences are in grayscale and some contain multiple subjects.

4.1.4.5 Hollywood human action dataset

A collection of actions in feature length movies has been introduced in [187]. This Hollywood human action (HOHA, Figure 4.1(e)) dataset contains 8 actions (answer phone, get out of car, handshake, hug person, kiss, sit down, sit up and stand up), performed by a variety of actors. In total, 32 movies have been used to gather these samples. There are two different training sets, and a test set. Each set contains around 230 samples. One of the training sets is automatically annotated using scripts of the

movies, the other is manually labeled. There is a huge variety of performance of the actions, both spatially and temporally. Occlusions, camera movements and dynamic backgrounds make this dataset challenging. Most of the samples are at the scale of the upper-body but some show the entire body, or are a close-up of the face.

4.1.4.6 Other datasets

There are a number of datasets available that address specific settings. For example, the ViHASi dataset [280] contains synthetically generated sequences of silhouettes. The dataset is useful to systematically investigate the performance of silhouette-based approaches, but does not provide a realistic setting. The HumanEva dataset [322] has been recorded to allow for objective evaluation of human motion analysis algorithms. While a number of actions is performed by four subjects, the number of sequences is rather small for the thorough evaluation of human action recognition approaches.

4.2 Image representation

In this section, we discuss the different image representations. Ideally, the features that are extracted from the image sequences should be able to generalize over small variations in person appearance, background, viewpoint and action execution. At the same time, the representations must be sufficiently rich to allow for robust classification of the action (see Section 4.3). The temporal aspect plays an important role in the performance of actions. Some of the image representations explicitly take into account the temporal dimension, others extract image features for each frame in the sequence individually. In this case, the temporal variations need to be dealt with in the classification step.

We divide image representations into two categories: holistic representations and patch-based representations. The former encodes the visual observation as a whole. Holistic, or global, representations are powerful since they encode much of the information. However, in general they require more preprocessing, such as localization, background subtraction or tracking. Also, they are more sensitive to viewpoint, noise and occlusions. When the domain allows for good control of these factors, holistic approaches usually perform well.

Patch-based, or local, representations describe the observation as a collection of independent patches. Such patches are often centered around spatio-temporal interest points. Since the number of interest points varies depending on the observation, a histogram of codewords is often used. This ensures that the feature vector has a fixed length, but the spatial and temporal information is discarded. Recently, there has been an emphasis on work that retains this information. In contrast to the holistic approach, patch-based approaches are less sensitive to noise and partial occlusion, and do not strictly require background subtraction or tracking.

We discuss holistic and patch-based image representations in Sections 4.2.1 and 4.2.2, respectively. A small number of works report the use of very specific features. We discuss these briefly in Section 4.2.3.

4.2.1 Holistic representations

Holistic representations regard the observation as a whole. Often, this requires localizing the person, which is the task of determining the region of interest (ROI) in the image. The observation within the ROI is subsequently encoded into a convenient image representation. Common global representations are derived from silhouettes, edges or optical flow, and we discuss these in Section 4.2.1.1. Such representations are global, and are therefore sensitive to noise, partial occlusions and variations in viewpoint. To partly overcome these issues, grid-based approaches spatially divide the observation into cells, each of which encodes the observation locally. Grid-based work is discussed in Section 4.2.1.2. Multiple images over time can be stacked, to form a 3-dimensional space-time volume, where time is the third dimension. Such volumes can be used for action recognition, and we present work in this area in Section 4.2.1.3.

4.2.1.1 Global representations

When information about the background is given, the silhouette of a person in the image can be obtained by using background subtraction. In general, these silhouettes contain some noise due to imperfect extraction. Moreover, they are sensitive to different viewpoints, and implicitly encode the anthropometry of the subject. However, they encode a great deal of information, and are insensitive to changes in appearance. When the silhouette is obtained, there are many different ways to encode either the silhouette area, or the contour.

One of the earliest uses of silhouettes is by Bobick and Davis [35]. They extract silhouettes from a single view, calculate differences between subsequent frames and aggregate these over all frames of an action sequence. This results in a binary motion energy image (MEI), which indicates where motion occurs. Also, a grayscale motion history image (MHI) is constructed, where pixel intensities are a recency function of the silhouette motion. Two templates are compared using Hu moments. Wang *et al*. [384] apply a $\Re$ transform to extracted silhouettes. This results in a translation and scale invariant representation, which is reduced in dimensionality using principal component analysis (PCA). Souvenir and Babbs [334] calculate a $\Re$ transform surface where the third dimension is time. Contours are used in [48], where the star skeleton describes the angles between a reference line, and the lines from the center to the gross extremities (head, feet, hands) of the contour. A codebook of star skeletons is used to compare sequences. Wang and Suter [374] use either a silhouette or a contour descriptor. Given a sequence of frames, an average silhouette is formed by calculating the mean intensity over all centered frames. Similarly, the mean shape is formed from the centered contours of all frames. Weinland *et al*. [386] match two silhouettes using Euclidean distance. In later work [385], it is shown that silhouette templates can also be matched against edges using Chamfer distance, thus eliminating the need for background subtraction.

When multiple cameras are employed, silhouettes can be obtained from each. Huang and Xu [139] use two orthogonally placed cameras at approximately similar height and distance to the subject. Silhouettes from both cameras are aligned at

the medial axis, and an envelope shape is calculated. Cherla *et al*. [50] also use orthogonally placed cameras and combine features of both. Such representations are somewhat view-invariant, but focus on protrusions on the human body, which are not always present. In the work by Weinland *et al*. [388], silhouettes from multiple cameras are combined into a 3D voxel model. Such a representation is informative but accurate calibration of the cameras is needed. They use motion history volumes (see Figure 4.2(b)), which is an extension of the MHI [35] to 3D. Matching is performed by first aligning the volumes using Fourier transforms on the cylindrical coordinate system around the medial axis. This makes the approach viewpoint-invariant.

Instead of silhouettes, the observation within the ROI can also be described with optical flow. This is the pixel-wise oriented difference between subsequent frames and can be seen as a motion descriptor. Flow information does not depend on the person's appearance and is somewhat independent of a person's pose. However, dynamic backgrounds can introduce noise in the motion descriptor. Also, camera movement results in observed motion, which is usually compensated by tracking the person. Efros *et al*. [73] calculate optical flow in person-centered images that are obtained from a tracker. They use sports footage, where persons in the image are very small. Optical flow can result in noisy displacement vectors, therefore the result is blurred. To make sure that oppositely directed vectors do not even out, the horizontal and vertical components are divided into positively and negatively directed, yielding 4 distinct channels. The similarity between two flow descriptors is measured using cross-correlation distance. Ahad *et al*. [8] use these four flow channels to solve the issue of self-occlusion when using a MHI approach. Ali and Shah [11] derive a number of kinematic features from the optical flow. These include divergence, vorticity, symmetry and gradient tensor features. In a subsequent step, PCA is applied to determine dominant kinematic modes.

4.2.1.2 Grid-based representation

Holistic representations are sensitive to noise, partial occlusions and changes in viewpoint due to a global matching function. These issues can be partly overcome by dividing the ROI into a fixed spatial or temporal grid. By summarizing the image observation within each cell in the grid and adapting the matching function, such a representation can be insensitive to small spatial and temporal variations. As each cell can be seen as a local descriptor, holistic grid-based representation somewhat resemble patch-based representations (see Section 4.2.2). However, in contrast to patch-based representations, holistic grid-based representation still require a global representation of the ROI.

Thurau [345] uses histograms of oriented gradients (HOG, [58]) and overcomes the high dimensionality by applying a codebook. In later work [346], non-negative matrix factorization is used to focus on foreground edges. Kellokumpu *et al*. [172] calculate local binary patterns along the temporal dimension and store a histogram of non-background responses in a spatial grid. Lu and Little [204] apply PCA after calculating the HOG descriptor over a grid, which greatly reduces the dimensionality. Instead of an unsupervised dimensionality reduction, Poppe and Poel [277] (see also Chapter 5) use silhouette gradients and learn a discriminative reduction between

pairs of classes. Ragheb *et al*. [279] first create a space-time volume (see also Section 4.2.1.3) by concatenating silhouettes over a given sequence. For each spatial location, the binary silhouette response is transformed into the frequency domain. A spatial grid is used, where each cell contains the mean frequency response of all spatial locations within the cell.

Optical flow can also be used in a grid-based representation. Danafar and Gheissari [59] adapt the work of Efros *et al*. [73] to a grid-based representation. Horizontal slices are used, that approximately divide the observation into head, body and legs. A similar approach was taken by Zhu *et al*. [413], who instead used vertical slices to distinguish between left and right tennis swings. Zhang *et al*. [408] use an adaptation of the shape context, where each log-polar bin corresponds to a histogram of motion word frequencies. Combinations of flow with shape descriptors are also common, and overcome the limitations of a single representation. İkizler *et al*. [142] combine the work of Efros *et al*. [73] with histograms of oriented line segments, obtained from an edge map. Flow, in combination with local binary patterns was used in [402]. Tran *et al*. [349] use grids of silhouettes and flow. Within each cell, a circular grid is used to accumulate the responses.

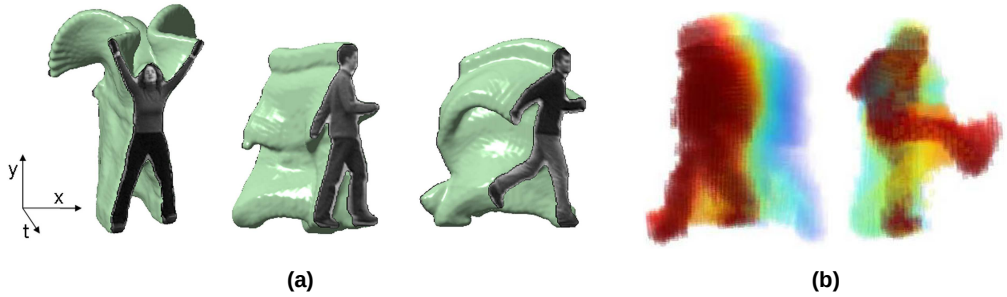
4.2.1.3 Space-time volumes

When frames over a given sequence are stacked together, a 3D spatio-temporal volume (STV) can be formed. Usually, frames are aligned to account for translation of the person in the image. In several of the works, the STV is sampled locally. While this approach shares many similarities with patch-based approaches, an STV is a holistic descriptor. The construction of an STV therefore requires accurate localization and alignment and, in many cases, background subtraction or tracking. This makes them less suitable for domains where patch-based approach typically perform well.

Blank *et al*. [30; 114] first stack silhouettes over a given sequence to form an STV (see Figure 4.2(a)). Then they use the solution of the Poisson equation to derive local space-time saliency and orientation features. Global features for a given temporal range are obtained by calculating weighted moments over the local features. Achard *et al*. [1] use a set of space-time volumes for each sequence, each of which covers only a part of the temporal dimension. Niyogi and Adelson [247] extract an STV by stacking frames, and apply spatio-temporal snakes to carve the volume. By analyzing the periodicity in the XT-slices at approximately knee height, different gait patterns, viewed from the side, are recognized.

Instead of a global matching, several works sample the STV surface and extract local descriptors. Yilmaz and Shah [405] use local differential geometric properties on the STV surface. Such properties include maxima and minima in the space-time domain. An action sketch is the set of descriptors that are found on the surface. Given that the descriptors are local, the method is sensitive to noise on the surface. The idea is extended by Yan *et al*. [401] by first constructing 3D exemplars from multiple views, for each frame in a training sequence. Then, for each view, an action sketch is calculated from the view-based STV and projected onto the constructed 3D exemplars. The action sketch descriptors encode both shape and motion, and can be matched with observations obtained from arbitrary viewpoints. Grundmann *et al*.

[117] extend the shape context to 3D and apply it to STVs. The sampling of interest points is adapted to give more importance to moving regions. Batra *et al.* [22] use silhouettes, and sample the volume with small 3D binary space-time patches. Oikonomopoulos *et al.* [251] extract salient points, and fit B-splines to these points to approximate an STV. The components of the partial derivatives of the volume are clustered into a codebook and used for training and recognition.



(a) (b)

Figure 4.2: (**a**) Space-time volume of stacked silhouettes (reprinted from [114], © IEEE, 2007) (**b**) Motion history volumes (reprinted from [388], © Elsevier, 2006). Even though the representations appear similar, (**a**) is viewed from a single camera, whereas (**b**) shows a recency function over reconstructed 3D voxel models.

Jiang and Martin [154] use motion and calculate shape flows over time. This results in flow lines in 3D, calculated at edge points. They propose a matching function that is invariant in scale and allows for small variations in space and time. Moreover, the matching can deal with cluttered backgrounds. Ke *et al.* [168] construct an STV of flow, and sample the horizontal and vertical components in space-time using a 3D variant of the rectangle features proposed in [365]. Ogata *et al.* [250] combine the work of [168] with that of Efros *et al.* [73]. A combination of STVs of silhouettes and flow is used by Ke *et al.* [170]. No background subtraction is needed, as 3D super-pixels are obtained from segmenting the STV. Action classification is cast as 3D object matching, where the distance to the segment boundary is used as a similarity measure. The work is extended in [169] to allow for the matching of parts, thus allowing recognition of actions under partial occlusion. The method can automatically segment actions in both space and time, but does so by sliding a temporal window through all 3 dimensions.

4.2.2 Patch-based representations

In many cases, it is not possible to robustly obtain a holistic representation. The lack of a background model, inaccurate localization and partial occlusions inhibit estimation of the ROI, or make the observation within the ROI unsuitable for further processing. In such cases, patch-based approaches can be used. These sample patches from the observation, which are used to form a final representation. Patches can be sampled densely over the ROI, or at specific points that have a high probability of corresponding to interesting motions. Motivated by their frequent use, we discuss these space-time interest points in more detail in Section 4.2.2.1.

A patch can be either 2D, or 3D. In their most simple form, all patches are treated independently and each action class is described by a distribution over all local patches. This bag-of-words approach is described in Section 4.2.2.2. Similar to holistic representations, observations can be grouped locally within a grid. This partly preserves the spatial or temporal information. We discuss several grid-based representations in Section 4.2.2.3. In many cases, there is a relation in space and time between the patches. By exploiting these correlations, actions can be modeled better since it allows suppression of noise to some extent. Also, the number of features required to model an action can be reduced. We discuss these correlations in Section 4.2.2.4.

4.2.2.1 Space-time interest points

Instead of dense sampling, patches are often extracted at space-time interest points. These points are likely to be distinctive, and often correspond to sudden changes in movement or appearance. Usually, points that undergo a translational motion in time will not result in the generation of space-time interest points. Several variants have been introduced.

Laptev and Lindeberg [186] extended the Harris corner detector [126] to 3D. Space-time interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. The scale of the neighborhood is automatically selected for space and time individually. The work is extended to compensate for relative camera motions in [185]. Oikonomopoulos *et al*. [252] extended the work on 2D salient point detection by Kadir and Brady [161] to 3D. In this approach, the entropy within each space-time patch (cuboid) is calculated, and the centers of those cuboids with local maximum energy are selected as salient points. The scale of each salient point is further determined by maximizing the entropy values.

One drawback of these methods is the relatively small number of stable interest points. Also, the approaches are unable to cope with subtle changes in space or time. These issues are partly addressed by Dollár *et al*. [70]. They use Gabor filtering on the spatial and temporal dimension individually. The interest points are obtained by selecting the local minima within a given neighborhood. Spatial and temporal scales of the neighborhood can be adjusted to influence the number of interest points that are selected. In a similar fashion, Rapantzikos *et al*. [286] apply a discrete wavelet transform in each of the three directions of a video volume. They obtain 8 different combinations by applying either a low-pass or a high-pass filter for each dimension. Each of these sub-bands corresponds to characteristic (slow, fast) movement in a certain dimension. The responses in the different sub-bands are used to select salient points in space and time. Oshin *et al*. [259] train randomized ferns to approximate the behavior of interest point detectors with lower computational complexity.

4.2.2.2 Bag-of-word representations

Patches are windows (2D) in an image or cuboids (3D) in a video volume. The number of patches depends on the size of the observation and on the density in which

they are extracted. In general, there can be many patches and the number is usually not fixed. This makes it harder to compare two sequences. Therefore, a codebook of patches is often used. A codebook is assembled by clustering patches and selecting either the cluster centra or the closest patch as a codeword.

Dense sampling is used by Scovanner *et al.* [308], who extend the SIFT descriptor [203] to 3D and construct histograms over codewords. In related work by Kläser *et al.* [175], 3D gradients are binned, for which regular polyhedrons are used. They extend the idea of integral images into 3D, which allows rapid dense sampling over multiple scales and locations in both space and time. A number of approaches employ a bank of filters and use the corresponding responses as the image representation. Chomat *et al.* [53] use spatio-temporal receptive fields and sample the ROI densely. They construct multi-dimensional histograms for the outputs of the filters. In related work by Jhuang *et al.* [152], several stages are used to ensure invariance to a number of factors. Their approach is motivated by the human visual system. At the lowest level, Gabor filters are applied to dense flow vectors, followed by a local max operation. Then the responses are converted to a higher level using stored prototypes and a global max operation is applied. A final matching stage with prototypes results in the final representation. The work in [241] is similar in concept, but uses different window settings. Schindler and Van Gool [306] extend the work by Jhuang *et al.* [152] by combining both shape and flow responses. Escobar and Kornprobst [84] used motion-sensitive responses and also consider interactions between cells, which allows them to model more complex properties such as motion contrasts.

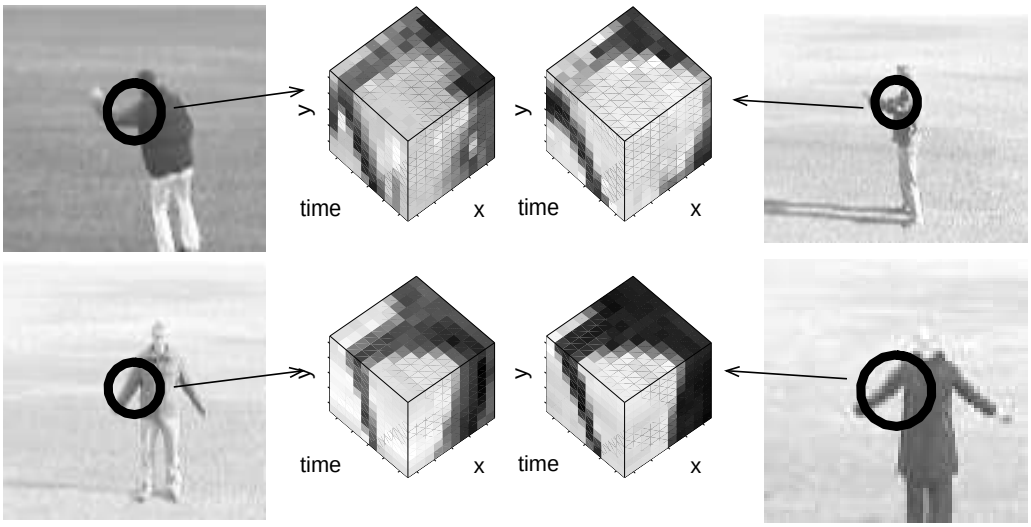


Figure 4.3: Extraction of space-time cuboids at interest points from similar actions performed by different subjects (reprinted from [185], © Elsevier, 2007).

Instead of dense sampling, patches can be selected only around interest points. This concept is visualized in Figure 4.3. For example, Schüldt *et al.* [307] calculate patches of normalized derivatives in space and time. A histogram of distances to codewords is used as a final representation. Niebles *et al.* [240] learn bags-of-words in an unsupervised fashion as an intermediary representation, and learn a distribution over

these representations to model each class. The words are concatenated brightness gradients, projected onto a lower dimension using PCA.

4.2.2.3 Grid-based representation

Similar to holistic approaches, described in Section 4.2.1.2, grids can be used to bin the patches spatially or temporally. Compared to the bag-of-words approach, using a grid ensures that spatial information is maintained to some degree. Moreover, a grid-based approach results in a significantly reduced descriptor size.

In the spatial domain, İkizler and Duygulu [144] extract oriented rectangular patches, which they bin into a grid. Each cell has an associated histogram that represents the distribution of rectangle orientations. They first apply background subtraction to reduce the number of false matches. However, the approach could be used without segmentation as well. In İkizler *et al*. [143], poses are first extracted using [281], and encoded as circular histograms of oriented rectangles. Action recognition is performed per frame. Zhao and Elgammal [411] bin interest points in a histogram with different levels of granularity. Interest points are weighted according to their temporal distance to the current frame.

Nowozin *et al*. [248] use a temporal instead of a spatial grid. The cells overlap, which allows them to overcome small variations in performance. Observations are described as PCA-reduced vectors around extracted interest points, mapped onto codebook indices.

Laptev and Pérez [188] bin histograms of oriented gradients and flow, extracted at interest points, into a spatio-temporal grid. This grid spans the volume that is determined based on the position and size of a detected head. The distribution of these histograms is determined for every spatio-temporal cell in the grid. Three different block types are used to form the new feature set. The plain type corresponds to a single cell, the temporal type is a concatenation of two temporally neighboring cells and the spatial variant combines several spatially neighboring cells. A subset of all possible blocks within the grid is selected using AdaBoost. Recent work by Laptev *et al*. [187] is similar in vein. Interest points are detected at multiple scales, and the neighborhood of each point is described as a histogram of oriented gradients and flow. Different grids are proposed, with varying numbers of spatial and temporal divisions and overlap settings. Flow descriptors from [73] are used by Fathi and Mori [88], who select a discriminative set of low-level flow features within space-time cells, which form an overlapping grid. In a subsequent step, a set of these mid-level features is selected to form the final classifier. Selection of both low-level and high-level features is performed between two classes, using the AdaBoost algorithm.

4.2.2.4 Correlations between features

In previous sections, distributions of words were used, either globally, or within a fixed grid representation. Such representations ignore the fact that there are often correlations between these words. Exploiting these correlations can lead to a reduced number of features. Moreover, by focussing on correlations between these words, one can improve classification between actions that have similar distributions of words,

but different co-occurrences between these words. Especially when using interest points, these might be generated due to noise in the background. Since this noise is often uncorrelated, focussing on correlations can reduce the influence of noise on the classification.

Scovanner *et al*. [308] construct a word co-occurrence matrix, and merge those words that have similar co-occurrences with other words. This step is applied several times, until the difference between all pairs of words is above a specified threshold. Similar in concept is the work by Liu *et al*. [200], who use a combination of the space-time features in [70], and spin images, which globally describe an STV extracted from silhouettes. A co-occurrence matrix of the features and the action videos is constructed. Then, the matrix is decomposed into eigenvectors and subsequently projected onto a lower-dimensional space. This embedding can be seen as a contextual feature-level fusion. Instead of determining pairs of correlated codewords, Patron-Perez and Reid [264] approximate the full joint distribution of features using first-order dependencies. Here, features are binary variables that indicate the presence of a video word. A maximum spanning tree is formed by analyzing a graph between all pairs of features. The work by Kim *et al*. [173] is different in the sense that correlation between two videos is measured. Canonical correlation analysis is extended to handle image sequences. The approach implicitly deals with affine variations. Discriminative features are subsequently selected using AdaBoost.

Song *et al*. [333] use a triangulated graphical model to detect and recognize human actions. Such models can be learned either supervised or unsupervised. Feature points are obtained in each frame, and tracked subsequently to establish correspondences. The location and velocity of each point are used as image cues. In Fanti *et al*. [85], additional local appearance cues are used. Moreover, global variables are introduced to represent properties such as scale, viewpoint and translation. Both works are evaluated on the detection and recognition task of directed walking motion. One drawback is that backgrounds are assumed to be static. Also, feature points are tracked between frames and camera motion, or rigid motion due to moving objects in the background, generates tracks that do not belong to the subject. This limitation is partly addressed by Niebles and Fei-Fei [238], who perform frame-based recognition by modeling the frame as a mixture of constellations. Each constellation models the spatial arrangement of video words, instead of tracked features. The video words are obtained from static and spatio-temporal features. Filipovych and Ribeiro [95] adapt the work in [238] to include both pose constellations and dynamics constellations. Star graphs of static and dynamic features are combined into a tree graph by conditioning on the landmark vertices of the individual graphs. Moreover, the models are trained without supervision.

Mikolajczyk and Uemura [219] extract for each frame a large number of local shape and motion features at edge locations. The relative location and orientation to a person's center of mass are stored with each feature, together with the annotated action label. These features are clustered and represented in a large number of vocabulary trees. By matching features extracted from an unseen frame, votes are cast over the person's location, orientation and action label. This approach allows them to simultaneously localize multiple persons in each frame. In Uemura *et al*. [355], global

motion patterns are detected and compensated for, in order to recognize action from moving cameras. In related work by Gilbert *et al*. [111], a large number of corners in xt, yt and xy planes are extracted first. Next, the relative spatial arrangement to all other corners is determined. This results in an extremely large number of features. Data mining techniques are further used to discriminatively select those combinations of features that are informative of a class. As such, they can also localize the action within the image.

Several works introduce hidden variables that correspond to action categories. Probabilistic latent semantic analysis (pLSA) is used by Niebles *et al*. [240]. The aim of pLSA is to learn the relation between sequences, codewords obtained from space-time interest points, and hidden action labels. Essentially, each action label corresponds to a distribution of codewords. These labels are learned in an unsupervised way from a collection of training sequences. Wong *et al*. [391] extend the pLSA model by including the location of a person's centroid. One drawback of these approaches is that the number of action labels needs to be determined empirically. Wang *et al*. [383] address this issue by taking a supervised approach. They introduce a semi-latent Dirichlet allocation (S-LDA) model, which is related to pLSA. Moreover, their motion features allow them to represent each frame by a single codeword, instead of a bag-of-words as used in [240]. In Wang and Mori [381], an adapted hidden conditional random field (hCRF) model is used to discriminatively learn constellations of local motion patches.

Savarese *et al*. [305] introduce the concept of space-time correlatons. These not only describe which words co-occur, but also look at the spatio-temporal relation between these words within a specified neighborhood. The performance of the final classifier is strongly dependent on the size of the codebook. Too few entries do not allow for good discrimination, while too great a codebook size is likely to introduce noise due to sparsity of the histograms. To overcome this issue, Liu and Shah [201] determine the optimal size of the codebook first using maximization of mutual information. This technique merges two codebook entries if they have comparable distributions given a set of image sequences. In addition, they apply correlograms and spatio-temporal pyramid matching. The correlograms are learned in a supervised fashion, and capture spatial correlations, while the spatio-temporal pyramid matching ensures that temporal information can also be exploited.

In above works, patches are extracted first, and subsequently the correlation between them is determined. In contrast, Wong and Cipolla [390] first detect subspaces of correlated movement. These subspaces correspond to large movements such as a waving arm. Within these spaces, a sparse set of interest points is detected, and different representation techniques are evaluated.

The effect of viewpoint on the recognition of human actions has received relatively little attention. Of note is the work of Farhadi and Tabrizi [86], who explicitly address the correlations between actions observed from different views. To this end, they use a split-based representation to describe clusters of codewords in each view. Given an action sequence that is observed from multiple views, the transfer of these splits between views can be determined.

4.2.3 Application-specific representations

The representations that were discussed in the previous sections are general in the sense that they can be used in a large number of application domains. When using feature selection techniques, the most suitable features or combinations can be selected. In contrast, a number of works use representations that are directly motivated by the application domain.

Joint locations or joint angles, either in 2D or 3D, are rich representations. The assumption is that these locations can be obtained in a pose recovery step (see also Chapter 6). See [97; 273] for an overview of literature in this domain. In 3D, the representations are completely view-invariant, whereas for 2D, there have been several approaches proposed to address the issue of matching 2D joint trajectories to action labels (e.g. [10; 262; 285; 313; 314; 404]). Since we focus on the recognition of human actions from image and video, we do not discuss these works here.

Smith *et al*. [332] use a number of specifically selected features. Some of these are low-level, and deal with color and movement. Also, higher-level features are used that are obtained from detected head and hand regions. A boosting scheme is used that takes into account the history of the action performance. The work by Vitaladevuni *et al*. [366] is inspired by the observation that human actions differ in accelerating and decelerating force. They identify reach, yank and throw types. Temporal segmentation into atomic movements, which are represented as ballistic words, is performed first. These include the movement type, spatial location with respect to a human center, and the direction of the movement.

4.3 Action classification

Given the image representation of an unseen sequence, the recognition of human action becomes the process of action classification. In this process, a label is associated to the observed image sequence. Alternatively, a probability distribution over the action labels can be given.

Traditionally, action classification is divided into template matching and state-space approaches (see e.g. [5; 373]). Recently, many different approaches have been proposed, and we feel that this traditional taxonomy does not capture these trends properly. Therefore, we use a different approach, that is more focussed on recent trends.

Section 4.3.1 discusses approaches that directly match new sequences to training sequences or action prototypes. These methods do not explicitly model variations in the temporal domain. A subcategory is that of discriminative classifiers that do not match, but rather classify the image representation directly. Grammars and graphical models are described in Section 4.3.2. These approaches have a state-space character, and model temporal variation implicitly. A topic that is related, but is strictly not within the scope of our survey is the detection of query motions in video. These approaches are useful to temporally (and spatially) divide a video into segments, but they lack both the action model and the labeling ability. We therefore discuss these works separately in Section 4.3.3.

4.3.1 Direct classification

This section discusses approaches that classify the image representation without paying special attention to variations in the temporal domain. In Section 4.3.1.1, we discuss work that maps a new sequence to labeled sequences in the training set or to action class prototypes. The traditional class of spatio-temporal templates also falls into this category. A second class of approach is that of the discriminative classifiers. These learn a function that discriminates between two or more classes by directly operating on the image representation. Approaches that use boosting schemes are also part of this category, which we discuss in Section 4.3.1.2.

4.3.1.1 Nearest neighbor classification

k-Nearest neighbor (NN) classifiers are the simplest methods of classification. The idea is that image representations of a given sequence are compared to those of labeled sequences in a training set. The most common label among the k most similar sequences is chosen as the classification. The ability to cope with variations in spatial and temporal performance, viewpoint and image appearance depends on the image representation that is used, and the distance metric that is applied.

NN classification can be either performed at the frame level, or for whole sequences. In the latter case, issues with different frame lengths need to be resolved. Due to their fixed descriptor length, holistic approaches lend themselves well for matching. For the patch-based representations, a histogram of codewords can be used to obtain a fixed-length descriptor. For example, Blank *et al*. [30] apply 1-NN using Euclidean distance between global features, Batra *et al*. [22] use Euclidean distance between histograms. Wang *et al*. [374] experiment with various distance metrics. Bobick and Davis [35] describe their MHI templates using Hu moments. Given the different orders of these moments, Mahalanobis distance is used to compare a given sequence to an action class prototype. Rosales' [297] work is related, but PCA is used to reduce the dimension. Tran *et al*. [349] use a learned discriminative distance metric in their NN classification.

When individual sequences are used for comparison, there is the risk that outliers will have a large impact on the final classification. Also, the computational performance of the nearest neighbor classifier is linear in the number of training sequences, which might cause problems when there are many of these sequences available. Instead, action class prototypes can be used with 1-NN classification. Prototypes can be obtained by simply averaging over sequences with similar class labels, as in [374]. Poppe and Poel [277] also take this approach, but learn discriminative feature transforms and distinguish between pairs of classes. Weinland *et al*. [388] create 3D voxel representations from multiple views. Mean dimensionality-reduced vectors are used as action class prototypes and compared with Mahalanobis distance. One drawback of action class prototypes is that they are not able to model more complex class distributions. This issue is addressed by Rodriguez *et al*. [291], who describe a method to generate a single spatio-temporal template that effectively captures the intra-class variance. The response of the filter is analyzed in the frequency domain, which makes the matching more effective.

Dynamic time warping Dynamic time warping (DTW) is a distance measure between two sequences, possibly with different lengths. It simultaneously takes into account a pair-wise distance between corresponding frames and the cost of alignment of the sequences. For two sequences to have a low alignment cost, they need to be segmented similarly in time, and be performed at similar rates. Dynamic programming is used to calculate the optimal alignment. Veeraraghavan *et al*. [364] use DTW but observe that their normalized shape features lie on a spherical manifold. Therefore, they adapt the distance function between two shapes. In later work [363], they also address the alignment of sequences by considering the space of warping functions for a given activity. A related distance is longest common subsequence (LCS), which is also applied between two sequences. It only takes into account similar elements of both sequences, and results in an increased distance when more inserts or deletions are necessary to warp one sequence onto the other. LCS has been used by Yang *et al*. [402].

Manifold comparison Different instances of a given action occupy only a part of the entire feature space. This subspace is a manifold, and it can often be embedded into a lower dimensional space. This embedding can be learned from training data, and allows for interpolation of the image representation. Elgammal and Lee [77] use this for human pose recovery, for which they construct manifolds for each action class, and learn mapping functions from image representation to manifold, and from manifold to pose space. For action recognition, pose information is not required. Instead, given a new sequence, the minimum distance of each frame to the manifold of a certain action can be determined. This approach has been taken by Masoud and Papanikolopoulos [210], who use PCA on motion recency images to determine the manifold. While the temporal order is neglected in such an approach, the burden of temporal alignment and variations in speed of performance can be overcome.

Instead of using PCA, which is linear, some works learn a non-linear embedding. Chin *et al*. [52] learn manifolds using either PCA or local linear embedding (LLE) on silhouette images. They experiment with different projection functions for LLE. Silhouettes and their distance transforms are also used by Wang and Suter [377] who use locality preserving projections (LPP) for the embedding. The use of Gaussian mixture models (GMM) to model the density of the low-dimensional embedding is investigated. Related work by the same authors [375] either uses the minimum mean frame-wise distance to the manifold, as in [210], or a frame-order preserving variant. Here, it is assumed that the time between two subsequent frames is equal for the entire sequence. More robust in this sense is the work by Blackburn and Ribeiro [29], who use an adaptation of DTW. This requires adding a time dimension into the embedding, for which they use Isomap. Recent work by Turaga *et al*. [352] focusses on parametric and non-parametric manifold density functions, and describes appropriate distance functions for Grassmann and Stiefel manifold embeddings. All these manifolds are learned in an unsupervised manner, which does not guarantee good discrimination between related classes. Jia and Yeung [153] address this issue by learning an embedding that is discriminative both in a spatial and temporal sense. They propose local spatio-temporal discriminant embedding (LSTDE), which maps

silhouettes of the same class close in the manifold, and model temporal relations in subspaces of the manifold.

Orrite-Uruñuela *et al*. [258] capture, for each action, the variation in viewpoint and temporal offset in a Kohonen self-organizing map (SOM). As such, they can project an unseen sequence to the different SOMs and simultaneously recover viewpoint and action class.

Keyframes It has been observed that many actions can be represented by a small number or even a single key frame or key pose. For example, Sullivan and Carlsson [336] recognize forehand and backhand tennis strokes by matching edge representations to stored and manually labeled key poses. Also based on edge distance is the work of Wang *et al*. [380], who learn action clusters in an unsupervised fashion. They manually provide action class labels after the clustering. Weinland *et al*. [387] also learn a set of action key poses but use 3D voxel representations.

The previous methods used only a single frame for action classification. This is convenient when the frame contains a key pose, but in general will generate many false matches. By considering a sequence of poses over time, ambiguities can be reduced. See also a discussion in Schindler and Van Gool [306]. Dedeoğlu *et al*. [62] use histograms of matches to manually selected key poses. The length of the histogram equals the number of key poses, and each bin contains the number of frames that best match the corresponding key pose. The histogram is normalized and 1-NN is used for classification. The work by Weinland and Boyer [385] is similar, but the minimum distance of each key pose to the frames in the sequences is used instead. Moreover, a small set of discriminative key poses is selected automatically. Key poses are often used in combination with hidden Markov models (HMM). We discuss these in Section 4.3.2.1.

Instead of direct classification, key poses can also be used to detect actions and, in a subsequent step, use a computationally more complex algorithm for classification. This technique, keyframe priming, is used by Laptev and Pérez [188]. Zhao and Elgammal [412] also detect keyframes first. They use the conditional entropy of the visual codewords in each frame as a measure. Only the features of these keyframes are used in a subsequent bag-of-words classifier.

4.3.1.2 Discriminative classifiers

Discriminative classifiers distinguish between classes without explicitly modeling each. The image representation is simply regarded as a feature vector. Support vector machines (SVM) are popular classifiers that learn a hyperplane in feature space that is described by a weighted combination of support vectors. SVMs have often been used in combination with patch-based representations, such as in [152; 185; 307]. Here, the image representation must be of fixed length, for example a histogram of codewords over a sequence of frames. SVMs can be trained efficiently. Relevance vector machines (RVM) can be regarded as the probabilistic variant of the SVM. An additional advantage is that RVM usually results in a sparser set of support vectors. They have been used for action recognition by Oikonomopoulos *et al*. [252].

Boosting In a boosting framework, a final strong classifier is represented by a set of weak classifiers. Usually, each weak classifier uses only a single dimension of the image representation. As such, it can be used as a discriminative feature selection process. Boosting is used in many works, either as a feature selection step, or as the actual classifier. The most common boosting approach is AdaBoost [102], which has been used in [88; 188; 250]. LPBoost, a variant of AdaBoost which yields sparser coefficients and is reported to converge faster, is used in [248]. Smith *et al*. [332] introduce a boosting variant that can use history information in the boosting scheme.

4.3.2 Graphical models

State-based models, or graphical models, are discussed in this section. They consist of states, connected by edges. These edges model probabilities between states, and between states and observations. For the task of action recognition, an observation corresponds to the image representation at a given frame. Usually, one model is trained per action class. In this case, states correspond to phases in the performance of the action. Graphical models are either generative or discriminative. While they share many characteristics, they are conceptually different. Generative models learn a joint distribution over both observations and action labels. They thus learn to model a certain action class. In fact, generative models can produce sequences of observations for a given action, hence the name. In contrast, discriminative models learn probabilities of the action classes, conditioned on the observations. As such, they do not model a class, but rather focus on differences between classes. We discuss generative and discriminative graphical models in Sections 4.3.2.1 and 4.3.2.2, respectively.

4.3.2.1 Generative graphical models

Hidden Markov models (HMM) are the most well-known generative graphical models. They use hidden states that correspond to different phases in the performance of an action. HMMs model state transition probabilities, and observation probabilities. To keep the modeling of the joint distribution over representation and labels tractable, two independence assumptions are introduced. First, state transitions are conditioned only on the current state, not on the state history. This is the Markov assumption. Second, observations are conditioned only on the current state, so subsequent observations are considered independent. We discuss the use of generative graphical models, in particular HMMs, for the task of action recognition.

HMMs have been used in a large number of works. Yamato *et al*. [400] cluster grid-based silhouette mesh features to form a compact codebook of observations. They train HMMs for the recognition of different tennis strokes. Training of an HMM can be done efficiently using the Baum-Welch algorithm. The Viterbi algorithm is used to determine the probability of observing a given sequence. When using a single HMM per action, action recognition becomes finding the action HMM that could generate the observed sequence with the highest probability. Per action, Niu and Abdel-Mottaleb [246] use a set of HMMs, each of which models the action from a certain viewpoint. Weinland *et al*. [386] construct a codebook by discriminatively selecting a set of templates. In their HMM, they explicitly include the viewpoint, which

allows them to condition the observation on the viewpoint. Related work by Lv and Nevatia [207] uses an Action Net, which is constructed by considering key poses and viewpoints. Transitions between views and poses are encoded explicitly. Ahmad and Lee [9] take into account multiple viewpoints and use a multi-dimensional HMM to deal with the different observations. Instead of modeling viewpoint, Lu and Little [204] use a hybrid HMM, where one process denotes the closest shape-motion template, while the other encodes position, velocity and scale of the person in the image. Ramanan and Forsyth [283] track persons in 2D by learning the appearance of the body-parts. In [282], these 2D tracks are subsequently lifted to 3D using stored snippets of annotated pose and motion. An HMM is used to infer the action from these labeled codeword motions. Feng and Perona [92] use a slightly different approach by assigning one codeword observation to each state. This allows them to effectively train the dynamics, at the cost of reduced flexibility due to a simpler observation model.

Instead of modeling the whole human body as a single observation, an HMM can be made for every body-part individually. This makes training easier, as the combinatorial complexity is reduced to learning dynamical models for each limb individually. In addition, this has the advantage that composite movements that are not in the training set can be recognized. İkizler and Forsyth [145] use the 3D body-part trajectories that are obtained using [282]. Instead of using labeled codeword motions, they construct HMMs for the legs and arms individually, where 3D trajectories are the observations. This allows them to use much simpler action models. For each limb, states of different action models with similar emission probabilities are linked. This results in a HMM that allows for automatic segmentation of actions, for legs and arms separately. A similar approach has been taken by Chakraborty *et al*. [46], where arms, legs and head are found with a set of view-dependent detectors. Lv and Nevatia [206] use a different approach, but they also use 3D joint locations as observations. First, they construct a large number of action HMMs, each of which uses a subset of the joints. This results in a large number of relatively weak classifiers. Subsequently, they use AdaBoost to select a set of these classifiers, that form the final strong classifier.

In the work by Peursum *et al*. [268], a factored-state hierarchical HMM (FS-HHMM) is used to jointly model image observations and body dynamics for each action class separately. By evaluating an image sequence using each of the action models, the action with the lowest log-likelihood is selected. Related work by Caillette *et al*. [44] uses a variable length Markov model (VLMM) to model observations and 3D poses for a given action. The work is mainly aimed at improved 3D pose tracking, but can also be used for recognition as in [268]. Natarajan and Nevatia [232] introduce a hierarchical variable transition HMM (HVT-HMM), which consists of three layers. The top layer models composite actions, the middle layer primitive actions and the bottom layer poses. Due to their variable window approach, actions can be recognized with low latency.

Related to generative graphical models are grammars. These specify explicitly in which order parts of an action can be observed. Hatun and Duygulu [128] and Fihl *et al*. [94] use the edit distance between sequences of codewords, which allows them to cope with small variations and differences in rate of movement. Ogale *et*

al. [249] construct a probabilistic context-free grammar that specifies which pose pairs can be observed. The viewpoint is explicitly encoded. Probabilities of pose pairs are learned from training data, while small viewpoint changes are allowed. Turaga *et al*. [353] model an action as a cascade of linear time invariant (LTI) dynamical models. In an unsupervised way, they simultaneously learn the dynamical models and temporally segment a sequence. Similar models are grouped into action prototypes. Finally, the cascade structure is formed by learning n-grams over the sequence of action prototypes. This cascade can be regarded as a grammar that describes the production rules for each action in terms of a sequence of action prototypes.

4.3.2.2 Discriminative graphical models

The independence assumptions in HMMs imply that observations in time are independent, which is often not the case. Therefore, discriminative graphical models have been proposed, that learn a conditional distribution of action labels, given the observations. These models can take into account multiple observations on different timescales. Consequently, they can be trained in such a way that they learn to discriminate between action classes, rather than learning to model each class individually, as in generative models. Therefore, discriminative models are suitable for classification of related actions that would easily be confused using a generative approach. In general, discriminative graphical models require many training sequences to robustly determine all parameters.

Conditional random fields (CRF) are commonly used discriminative models that can model multiple overlapping features. Sminchisescu *et al*. [328] use a linear chain CRF, where the state dependency is first-order. They compare CRFs with generative HMMs and maximum entropy Markov models (MEMM). The latter are related to the CRF, but are directed models instead. They suffer from the label bias problem, in which states with few outgoing state transitions are favored. A more detailed comparison between CRFs and MEMMs is given in [183]. Sminchisescu *et al*. evaluated the performance of their CRFs using different observation window sizes. They show that CRFs outperform both MEMMs and HMMs when using larger windows. These results are partly supported by Mendoza and Pérez de la Blanca [212], who obtain better results for CRFs compared to HMMs using shape context features, especially for related actions (e.g. walking and jogging). Interestingly, when using motion features, HMMs outperformed CRFs.

Variants of CRFs have also been proposed. For example, Wang and Suter [376] use a factorial CRF (FCRF), which is a generalization of the CRF. Structure and parameters are repeated over a sequence of state vectors, which can be regarded as a distributed state representation. This allows for the modeling of more complex interactions between labels and long-range dependencies, while inference is approximate instead of exact as in CRFs. The authors report improved results for the FCRFs, compared to both linear-chain CRFs and HMMs. Natarajan and Nevatia [233] use a 2-layer graphical model, where the top level encodes action and viewpoint. On the lower level, CRFs are used to encode the action and viewpoint-specific pose observation. Ning *et al*. [245] combine a discriminative pose recovery approach with a CRF for action recognition. The parameters of both layers are jointly optimized. No

image-to-pose data is required during training, but has been shown to improve performance. Shi *et al*. [315] use a semi-Markov model (SMM), which is suitable for both action segmentation and action recognition. Properties that relate to segment boundaries, encode characteristics about segments, or capture interactions between neighboring segments are included in the feature representation.

4.3.3 Video correlation

There is a category of approaches that do not explicitly model the image representation of subjects in the image, nor do they model action dynamics. Rather, they correlate an unseen sequence to video sequences in a database. Such work is mostly aimed at the detection of actions, rather than their recognition. However, since these works share many similarities to those previously discussed, we will describe them briefly in this section. The detection of cyclic actions is discussed in Section 4.3.3.1.

Zelnik-Manor and Irani [406] use histograms of appearance-normalized gradient patches, calculated at multiple temporal scales. Patches that exhibit low variance in the temporal dimension are ignored, which focusses the representation on the moving areas in the video. Consequently, for human action recognition, this restricts the approach to detection of movement against non-moving backgrounds. Ning *et al*. [241] use histograms of codewords, obtained from Gabor response instead of gradient patch histograms.

Shechtman and Irani [312] consider the spatial dimension by correlating space-time patches over different locations in space and time. Similarly to [406], they use space-time cuboids, but local motion information is used instead of gradients. To avoid calculating the optical flow, a rank-based constraint is used directly on the intensity information of the cuboids. Matikainen *et al*. [211] present an approximation of method that uses motion words and a look-up table to allow for faster correlation of the motion of different patches. In recent work by Shechtman and Irani, [311], a self-similarity descriptor is proposed, that correlates local patches. Such a descriptor is invariant to color, texture and can deal with small spatial variations. A query template is described by an ensemble of all descriptors, either at the frame level, or over a sequence of frames. Junejo *et al*. [160] focus on detection of similar actions from multiple viewpoints. The key idea is to look at temporal similarities between the frames of a sequence. By observing the self-similarity matrix, actions seen from different viewpoints show remarkable resemblances (see Figure 4.4). A local approach is taken, where points on the diagonal are encoded as log-polar histograms. A sequence is described as a bag-of-features.

The approach by Boiman and Irani [36] is slightly different. They describe a sequence as an ensemble of local patches, which can be either spatial or spatio-temporal. A similarity score is based on the composition of a query sequence from the local patches. Similar sequences require less, but larger, patches compared to dissimilar sequences.

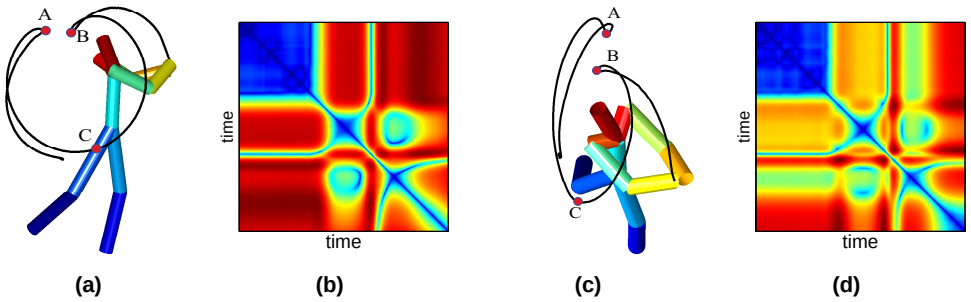


Figure 4.4: Example of cross-correlation between viewpoints. (**a**) and (**c**) show a golf swing seen from two different viewpoints. (**b**) and (**d**) show their corresponding self-similarity matrices. Note the similarity in structure (reprinted from [160], © Springer-Verlag, 2008).

4.3.3.1 Cyclic actions

Some works assume periodicity of the motion, which allows for segmentation by analyzing the self-similarity matrix. For example, Seitz and Dyer [309] introduce a periodicity detection algorithm that is able to cope with small variations in the temporal extent of a motion. They track markers and use an affine-invariant distance function, which makes the work invariant to changes in view, translation and scale. Cutler and Davis [56] use the video data directly to perform a frequency transform on the self-similarity matrix of a tracked object. Peaks in the spectrum correspond to the frequency of the motion. The type of action is determined by analyzing the matrix' lattice structure. Polana and Nelson [271] also use Fourier transforms to find the periodicity and temporally segment the video. They use motion features, which they match to known 2D motion templates.

4.4 Discussion

In recent years, the domain of human action recognition has seen tremendous progress. In this section, we point out limitations of the state of the art and and identify directions for future research to address these limitations.

Holistic image representations have proven to yield good results, and they can usually be extracted with low cost. However, their applicability is limited to scenarios where ROIs can be determined reliably. Moreover, they cannot deal with occlusions. Patch-based representations have been proposed to address these issues. Initial work focussed on bag-of-feature approaches but currently, more advanced representations are being used. These take into account correlations between patches, both spatially and temporally. Patch-based representations do not require precise determination of the ROI, but some knowledge about the location and scale of the person in the image is needed. One important issue that has been largely ignored is how to deal with more severe occlusions.

Most of the work described in this overview is restricted to fixed viewpoints. This limits the applicability to domains where movement is observed from similar view-

points as during training. The use of multiple view-dependent action models solves this issue, but at the cost of increased training complexity. Recently, transfer learning has been used to learn correspondences between views [86]. Such an approach decreases the dependence on training data from all viewpoints.

Regarding classification, we discussed direct classification and graphical models. In the former, temporal variations are not explicitly modeled, which has proved to be a reasonable approach in many cases. For more complex motions, it is questionable whether these direct measures suffice. Generative graphical models such as HMMs can model temporal variations but these have the disadvantage that distinguishing between related actions is difficult. In this respect, discriminative graphical approaches are more suitable. However, these models require a large amount of training data. Future work should therefore address the learning of action classifiers from only a few examples. A related issue is the flexibility of the classifier with respect to adding or removing action classes from the repertoire.

Currently, many approaches assume that the video is readily segmented into sequences that contain exactly one action instance. The detection task is thus ignored, which limits the applicability to situations where the segmentation of different actions is possible. Especially for applications that require real-time processing, reliable segmentation might prove a difficult task. In future work, the temporal segmentation of actions should therefore be addressed explicitly.

As discussed before, human action recognition is related to many other fields of research. It is to be expected that improvements over the current state of the art can be obtained by combining existing work with recent advances in human pose recovery. The work by [268; 328] shows that such work leads to good results, both for the recognition of actions, as for the recovery of human poses. Currently, the localization of the person in the image is often regarded as a preprocessing step. Since good localization is of key importance for reliable action recognition, it makes sense to combine these two tasks (e.g. [346; 355]).

Another important aspect of human action recognition is the current evaluation practice. The introduction of publicly available datasets (see Section 4.1.4) has greatly shaped the domain. They provide common training and test data which allow for objective comparison between different approaches. Moreover, they allow for better understanding of different methods since researchers are aware of the challenges of each dataset. However, the drawback of such sets is that algorithms may be biased to the dataset. This may lead to complex approaches, that perform slightly better on a given dataset, but may prove to be less generally applicable. Also given the increasing level of sophistication of action recognition algorithms, the introduction of more advanced and more elaborate datasets might be useful to focus research efforts on more realistic problems. For example, the HOHA dataset, first used in [187] moves the focus from controlled recordings to feature films where scripts and subtitles can be used for automatic annotation [55]. Such a shift raises the question as to what the application of our work is. Action recognition is used in surveillance, human-computer interaction and video retrieval, but these areas differ in many ways. Human-computer interaction applications require real-time processing, missed detections in surveillance are unacceptable and video retrieval applications often cannot

benefit from a controlled setting and require a query interface (e.g. [337]). Given these differences, it seems reasonable to record different datasets for the various domains. This would keep recording settings realistic, while focussing only on relevant action classes. Moreover, the use of application-specific datasets allows for the use of evaluation metrics that go beyond precision and recall measures, such as speed of processing or detection accuracy. Still, the compilation or recording of a dataset that contains sufficient variation in movements, recording settings and environmental settings remains challenging and should be a topic of discussion within the research community.

Given the current state of the art, and motivated by the broad range of applications that can benefit from robust human action recognition, it is to be expected that many of the previously mentioned challenges will be addressed in the near future. This would be a big step towards the fulfillment of the longstanding promise of the field to achieve robust automatic recognition and interpretation of human action.

5

Human action recognition using common spatial patterns

In the previous chapter, we presented an overview of vision-based human action recognition literature. In this section, we focus on recognition of human actions from a single view. Secondly, we require our approach to have low computational complexity, preferably to work in real-time. This requirement influences both the choice of image representation, and the classification approach.

We introduce an approach to human action recognition where we learn functions that discriminate between two classes. Yet, we avoid having to estimate a large number of parameters by representing actions as single prototypes. These prototypes lie in a space that is transformed by applying common spatial patterns (CSP) on the feature data. CSP is a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. After applying CSP, the first components of the transformed feature space contain high temporal variance for one class, and low variance for the other. This effect is opposite for the last components. For an unseen sequence, we calculate the variance over time, using only a fraction (the first and last components) of the transformed space. Each action is represented by the mean of the histograms of all corresponding training sequences, which is a very compact but somewhat naive representation. A simple classifier distinguishes between the two classes. All discriminant functions are evaluated pairwise to find the most likely action class. Even though such an approach inherently generates much noise in the classification, we show that we can accurately recognize actions, even when few training sequences are used.

The advantage of our method is that we require relatively few training samples. Despite considerable variation in action performance between subjects, we obtain reasonable results when training on data of a single subject. Also, we avoid retraining all functions when adding a new class, since the discriminative functions are learned pairwise, instead of jointly over all classes. Moreover, both feature extraction and classification can be performed in limited time. In fact, our approach can work in real-time.

For fully automatic recognition of human actions, we need to localize the humans

in the image first. Here, we assume that this localization has been performed previously, for example using work by Thurau [345] and Zhu *et al*. [414]. We use a grid-based silhouette descriptor, where each cell is a histogram of oriented silhouette gradients (HOSG). This representation resembles the concept of histograms of oriented gradients (HOG, [58]) and has been previously introduced in Section 3.2.4.1.

We discuss common spatial patterns, and the construction of the CSP classifiers, in Section 5.1. Our approach is evaluated on the Weizmann human action dataset. We summarize our extensive experiments in Sections 5.3 and 5.4. Discussion of our results, and a comparison with previous work appears in Section 5.5. A preliminary version of this section appeared in [277].

5.1 Common spatial patterns

Common Spatial Patterns (CSP) is a spatial filter technique that is often used in classifying brain signals [228]. It transforms temporal feature data by using differences in the variance between two classes. After applying the CSP, the first components of the transformed data have high temporal variance for one class and low temporal variance for the other. For the last data components, this effect is opposite. When transforming the feature data of an unseen sequence, the temporal variance in the first and last components can be used to discriminate between the two classes.

Consider the case where we have training sequences for two actions, a and b. Each training sequence can be seen as $n \times m_p$ matrix, where n is the number of features and m_p is number of time samples. We assume that the data is normalized in such a way that the mean of each feature is 0. Let C_a be the concatenation of the examples of action a, C_a is an $n \times m_a$ matrix. We do the same for action b to construct the matrix C_b. Now consider the matrix:

$$C = C_a C_a^T + C_b C_b^T \tag{5.1}$$

C is the variance of the union of the two data sets. Since C is symmetric, there exists a orthogonal linear transformation U such that $\Lambda = UCU^T$, a positive diagonal matrix. The next step is to apply the whitening transformation $\Psi = \sqrt{\Lambda}^{-1}$, which gives us $(\Psi U)C(\Psi U)^T = I$, and thus:

$$\begin{aligned} S_a &= (\Psi U) C_a C_a^T (\Psi U)^T && (5.2)\\ S_b &= (\Psi U) C_b C_b^T (\Psi U)^T && (5.3)\\ S_a + S_b &= I && (5.4) \end{aligned}$$

Since S_a is symmetric, there is an orthogonal transformation D such that DS_aD^T is a diagonal matrix with decreasing eigenvalues on the diagonal. Hence, DS_bD^T is also a diagonal matrix but with increasing eigenvalues on the diagonal. The CSP is the spatial transform $W = D\Psi U$ which transforms a data sequence into a sequence of dimension $2k$ such that a vector belonging to one action has high values in the first k components. For a vector of the other action, the situation is opposite. Hence, the temporal variance in these first and last components can be used to discriminate between action a and b.

5.1.1 CSP classifiers

Based on the CSP technique, we design discriminating functions $g_{a,b}$ for every action a and b with $a \neq b$. First we calculate the CSP transformation $W_{a,b}$ as described above. Then we apply $W_{a,b}$ to each action sequence of class a and b. Afterwards, the mean is taken over the entire sequence. This results in a single n-dimensional vector which can be considered a histogram, normalized for the length of the sequence. Next, we calculate the means $\bar{a}$ and $\bar{b}$ of these training vectors for action a and b, respectively. In order to compute $g_{a,b}(x)$ for an unseen action sequence x, we use the same procedure and first apply $W_{a,b}$ to x. We then calculate the variance over time over all components, which gives a vector x' of length n. Finally, $g_{a,b}(x)$ is defined as follows:

$$g_{a,b}(x) = \frac{\|\bar{b} - x'\| - \|\bar{a} - x'\|}{\|\bar{b} - x'\| + \|\bar{a} - x'\|} \tag{5.5}$$

Here, $\|x\|$ denotes the vector length, or norm, of x. Evaluation of a discriminant function gives a continuous output in the $[-1, 1]$ interval. Note that $g_{a,b} + g_{b,a} = 0$. With a rescaling and transform into the $[0, 1]$ domain, we could interpret these outputs as probabilities. However, since we assume equal prior probabilities for each class, we use our voting scheme for clarity. Also, we could have used different discriminative functions than Equation 5.5. For example, we could have kept the individual training vectors, instead of the mean. This would allow to better model intra-class variance. In this case, one could use Mahalanobis distance, or use a margin classifier such as Support Vector Machine (SVM). These alternatives are, however, sensitive to outliers in the data.

We combine our pairwise classifiers into a multi-class classifier using voting. Such a scheme has been proposed by Friedman [103] for binary outputs, i.e. $g'_{a,b}(x) = sgn(g_{a,b})$. We apply their work for continuous outputs, without loss of generality. In such a scheme, an action sequence is classified by evaluating all discriminant functions between pairs of a and b over all actions, and summing their votes:

$$g_a(x) = \sum_{a \neq b} g_{a,b}(x) \tag{5.6}$$

Since each action class appears in the exact same number of discriminative functions, the classification of x is the action η for which $g_a(x)$ is maximal. This is the class that receives most of the voting mass:

$$\eta(x) = \arg\max_a g_a(x) \tag{5.7}$$

Note that we also evaluate the discriminant functions in which the actual class does not appear. This introduces a large component of noise into the voting. However, actions that show more similarities with the unseen sequence will receive more mass in the voting. Hastie and Tibshirani [127] remark that such a voting approach tends to favor classes that are closer to the average value in feature space. Such an effect would be larger for weaker pairwise discriminative functions. In our experiments, the

dimensionality is relatively high compared to the number of classes and we expect that the effect of this bias is small.

More complex classification schemes are also possible. For example, Hastie and Tibshirani [127] take into account all individual pairwise probability estimates and minimize a Kullback-Leibler criterion to find the optimal decision boundaries. The advantage is that the decision boundaries are determined jointly for all pairs of classes. Works by Allwein *et al*. [12] and Dietterich and Bakiri [67] use error-correcting codes, where each 'bit' in the code corresponds to a pairwise decision. While these approaches are better at dealing with noise caused by incidental erroneous decisions, their added value in performance over voting is limited [12]. Moreover, we prefer the straightforward interpretation of the voting outcome.

Our multi-class classifier requires $m(m-1)/2$ functions, with m being the number of classes. Note that we could alternatively have used a one-vs-all classification scheme. In this case, we would have needed to learn a discriminative function for each class. While the complexity of such an approach is linear in the number of classes, instead of quadratic as in our scheme, the discriminative functions need to be more complex.

5.2 HOSG silhouette descriptors

Similar to our work on human pose recovery in Section 3.2, we use a grid-based approach. For action recognition, grids were used as an image representation by [144; 345; 376], see also Section 4.2.1.2. Our image representation (HOSG-*R*) is a variant of histogram of oriented gradients (HOG, [58]) and is similar to the one used in Section 3.2.4.1. To make this chapter self-containing, we summarize the processing steps used to obtain the descriptor. Subsequently, we explain differences between the descriptor used for human pose recovery (HOG-*F*, see Section 3.2.1).

The different steps in our approach are shown in Figure 5.1. Given an extracted silhouette, we determine the minimum enclosing bounding box, which determines the region of interest (ROI). We add space to make sure the height is 2.5 times the width. Next, we divide the ROI into a grid of 4×4 cells. Within each cell, we calculate the distribution of silhouette gradients. Since the gradient of binary silhouettes can only be vertical, horizontal or diagonal, we use 8 bins of 45° directed gradients instead of 20° undirected ones. Pixels that are not on silhouette boundaries are ignored.

This idea is similar to that of histogram of oriented gradients (HOG) but our implementation is a simplification at a number of levels. First, we do not apply a Gaussian filter to enhance the edges. Second, we do not use overlapping cells, which significantly reduces the size of our descriptor. Third, and most important, we only take into account the silhouette outline thus discarding the internal edges. The final 128-dimensional descriptor is a concatenation of the histograms of all cells, normalized to unit length to accommodate variations in scale.

Compared to the HOGs that we used for human pose recovery (HOG-*F*-5×6, see Section 3.2.1), we do not use the *fit* ROI, but the ROI where the *ratio* between height and width is 2.5. Also, we use a 4×4 grid instead of a 5×6 grid. This choice is motivated by the smaller ROIs that we use here. A grid with more cells

would be less robust to noise due to incorrect segmentation, small changes in ROI and inter-personal differences. Also, in human action recognition, we are interested in larger-scale movements. Another difference is that we use silhouette gradients instead of edge gradients. We will therefore refer to this representation as histograms of oriented silhouette gradients (HOSG-*R*). Specifically, we will use the HOSG-*R*-4×4 setting in this chapter. However, we will report the performance of our algorithm on different HOG-*R* and HOSG-*R* settings in Section 5.4.1.

Due to the normalization of the descriptor to unit length, and the relatively high dimensionality compared to the number of data points in a sequence, the covariance over a sequence may be nearly singular in some cases. We avoid this by applying PCA [158] and select the 50 first components. These explain approximately 75% of the variance, depending on the subject that is left out. See the next section for details regarding this process.

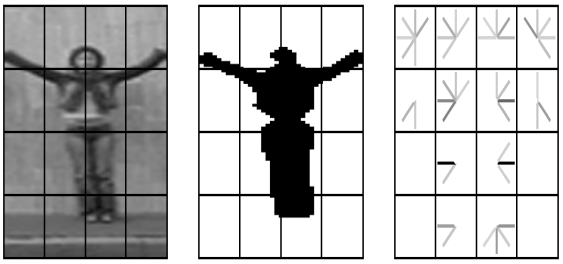


Figure 5.1: Silhouette descriptor, (**a**) image, (**b**) mask and (**c**) the boundary orientations, spatially binned into cells. Normal vectors are shown for clarity.

5.3 Experiment results

We evaluated our approach on a publicly available dataset which is briefly described in Section 5.3.1. We present the setup of our experiments and our obtained results in Sections 5.3.2 and 5.3.3, respectively. Additional experiments are described in Section 5.4. A discussion of the results and a comparison with related work are given in Section 5.5.

5.3.1 Weizmann human action dataset

For the evaluation of our approach, we used the Weizmann human action dataset [30; 114], see also Section 4.1.4.2. This set consists of 10 different actions, each performed by 9 different subjects (see also Figure 5.2). For subject *Lena*, additional sequences appear for the run, skip and walk action. We decided to leave these out in order to obtain a balanced set. This also allowed for direct comparison of our results to those previously reported on the dataset. Note that our approach also works for unbalanced sets. The skip action was not originally present in the set and we present results both with and without the skip action.

Figure 5.2: Example frames from the Weizmann human action dataset. Different subjects performing the actions bend, jack, jump, pjump, run, side, skip, walk, wave1 and wave2.

Each sequence is approximately 2.5 seconds long. There is considerable intra-class variation due to different performances of the same action by different subjects. Most notably, the jump, run, side, skip and walk actions were performed either from left to right, or in the opposite direction. Since the actions were performed on a slight slope, the direction of movement also results in slightly different movement style. Despite these differences, we treated performances in different directions as belonging to the same class. The sequences were recorded from a single camera view, against a static background, with minimal lighting differences. Binary silhouette masks are provided with the dataset. There is a considerate amount of noise in these silhouettes due to inaccurate background segmentation (see also Figure 5.6).

5.3.2 Experiment setup

We evaluated our method using leave-one-out cross-validation (LOOCV), where each of the 9 folds corresponds to all sequences of the corresponding subject. Specifically, this gave us 80 training sequences per fold, 8 for each of the 10 actions. First, we calculated the PCA transformation over all training sequences and projected the silhouette descriptors down onto the first 50 components. Next, we learned all discriminant functions $g_{a,b}$, between all pairs of actions a and b ($1 \leq a, b \leq 10, a \neq b$). Specifically, we used the first and last $k = 5$ components in the transformation, which gave us action prototypes vectors of dimension 10. We experimented with other values for k but found no improvement for $k > 5$. For each of the sequences of the subject whose sequences were left out, we evaluated all discriminant functions. Each of these evaluations softly votes over class a and b. In our final classification, we selected the class that received the highest voting mass.

5.3.3 Results

We performed the LOOCV experiment and obtained a performance of 95.56%. In total, 4 sequences were misclassified. The skip action of subject *Daria* was classified as jumping (2), the skip action of subject *Ido* was classified as running (3). Also, the jump action of subject *Eli* and the run action of subject *Shahar* were both classified

as walking (4 and 2, respectively). The number between brackets is the order of the correct label amongst the guessed labels. The confusion matrix for this experiment is shown in Figure 5.1 (left).

In order to be able to compare our results with those reported in previous studies, we also left out the skip class. This resulted in a performance of 96.30%. Again, the jump action of subject *Eli* and the run action of subject *Shahar* were classified as walking (4, 2). In addition, the wave1 action of subject *Lyova* was misclassified as wave2 (2).

In line with Friedman [103], we also evaluated the performance when using binary outputs for the discriminative functions (i.e. $g'_{a,b}(x) = sgn(g_{a,b})$). With the skip action, 3 additional errors were made which resulted in a performance of 92.22%. Without skip, the performance was similar to the soft vote case at 96.30%.

	Guessed									
Actual	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9									
jack		9								
jump			8					1		
pjump				9						
run					8			1		
side						9				
skip			1		1		7			
walk								9		
wave1									9	
wave2										9

	Guessed									
Actual	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	7	1							1	
jack		9								
jump		1	5			1	1		1	
pjump				9						
run					6			2	1	
side						9				
skip		1			3		2	3		
walk					1		1	7		
wave1									8	1
wave2		1								8

Table 5.1: Confusion matrices for Weizmann human action dataset including skip action with CSP (left, performance 95.56%), and without CSP (right, performance 77.78%). See text for explanation.

Both the feature representation and the classifier had an important impact on the performance. To measure the added value of using CSP, we performed an additional experiment where we did not transform the feature space. Instead, we took the first 10 components of the PCA. For each training sequence, we calculated the histogram by taking the mean of the feature vector over time which resulted in a 10-dimensional vector. We determined the prototype for each action by averaging these histograms. Again, we used Equations 5.5 and 5.7 to determine the class estimate. We achieved a performance of 77.78% for all actions, and 85.19% with the skip action omitted. The confusion matrix for all 10 actions is shown in Figure 5.1 (right). When we used the first 50 PCA components, the performance slightly increased to 80.00% for all actions, while the performance without the skip action remained the same. A closer look at the misclassifications shows confusion between run, skip and walk, along with some incidental confusions. It thus becomes clear that the use of CSP is advantageous over a feature representation without CSP transform.

The baseline for the full dataset is 10.00%, and 11.11% when the skip action is left out. Obviously, our results are well above these baselines and show that we can achieve good recognition, even when single action prototypes of dimension 10 are used. Also, it shows that intra-class variations can be handled without model-

ing the variance between different subjects. To gain insight in the characteristics of our method, we conducted additional experiments. These are described in the next section.

5.3.3.1 Computational performance

In this section, we present the computational performance of our approach, both during training and in classification. We used un-optimized Matlab code, which was evaluated on a Pentium IV 2.8 GHz computer. The same settings as in Section 5.3.2 were used, and the reported computation time is an average over all subjects using LOOCV. Since we used the foreground masks provided with the dataset, we cannot give the computation time required to calculate them. Given these masks, calculation of the HOG descriptor takes 2.6 *ms* per frame on average. The difference between the average time to calculate the HOG descriptor in Section 3.2.3.5 is mainly due to the large difference in image size, and the lower number of cells in the HOSG-*R*-4×4 descriptor. Also, given the use of silhouettes, no edge orientation binning is needed which further reduces computation cost. Learning the discriminative functions for all 10 actions using the training data of the remaining 8 subjects took on average 886.1 *ms*. This training step only needs to be performed once but it shows the computational simplicity of our approach. The average time to classify a test sequence was 12.5 *ms*. Given the average sequence length of 2.5 seconds, we can conclude that our approach works in real-time.

5.4 Additional experiments and results

In addition to the evaluations described above, we conducted several experiments to see how our approach performs with different settings and under different conditions. We used our HOSG-*R*-4×4 descriptors with the settings as described in Section 5.2), unless stated otherwise. Also, we used the standard Weizmann human action dataset, except in Section 5.4.3.

In Section 5.4.1, we use different image representations. Section 5.4.2 describes our experiments where we used only part of the available training data. Evaluations on sequences with different deformations and viewpoints are discussed in Section 5.4.3. Finally, describe our experiments with recognition from a small number of frames in Section 5.4.4.

5.4.1 Results using different image representations

In this section, we evaluate the effect of descriptor size and type on the classification performance. We used HOG-*R* and HOSG-*R* descriptors that are extracted within a ROI with fixed height-width ratio of 2.5. The former uses edges, extracted within a silhouette mask (see also Section 3.2.4.1). The latter is described in Section 5.2. We also used three different grid sizes: 3×3, 4×4 and 5×6. Note that HOSG-*R*-4×4 was used in the previous section. In addition, we evaluated the performance of the HOG-*F*-5×6 that was previously used in Section 3. This descriptor is similar to the HOG-*R*-4×4 but the ROI is the minimum enclosing rectangle of the foreground mask.

Descriptor sizes for HOG-*R* are 81, 144 and 270 for the three grid sizes respectively. For HOSG-*R*, these sizes are 72, 128 and 240. For HOG-*F*-5×6, the length is also 270. We kept the number of CSP components constant. Unseen sequences and action prototypes were both points in 10-dimensional space ($k = 5$).

We used the LOOCV approach for evaluation, with the data of 8 subjects for training and the data of the remaining subject for testing. The results are summarized in Table 5.2. HOSG-*R* performed slightly better than HOG-*R*. We can clearly see that 4×4 outperformed both smaller and bigger grids. We expect that 3×3 grids do not capture enough detail to distinguish between classes. For 5×6 grids, we contribute the lower performance to the smaller cell sizes. This causes the histograms to become sparse which results in higher similarity scores when small variations between performances of an action occur. In also appeared that the differences between the two ROI settings are small.

	HOSG-*R*			HOG-*R*			HOG-*F*
	3×3	4×4	5×6	3×3	4×4	5×6	5×6
All actions	84.44	95.56	85.56	83.33	90.00	87.78	85.56
Skip omitted	88.89	96.30	91.36	90.12	92.59	90.12	90.12

Table 5.2: Classification performance (in percent) using HOSG-*R*, HOG-*R* and HOG-*F* for different grid sizes.

5.4.2 Results using less training data

The fact that our approach is able to perform well, even though intra-class variation is not modeled, gives the impression that we can train our classifiers with less training data. Note here that training for all actions and all training subjects takes well under 1 second. To verify this hypothesis, we evaluate the performance of our approach using different numbers of subjects in the training set. Again, we used the LOOCV scheme. For each number of training subjects k, we present the results as averages over all $8!/(k!(8-k)!)$ combinations of training subjects. Table 5.3 summarizes these results, both using all actions, and with the skip action omitted.

Subjects	Combinations	All actions	Skip omitted
1	8	64.72%	69.14%
2	28	77.82%	83.82%
3	56	81.83%	88.98%
4	70	84.60%	90.85%
5	56	86.63%	92.44%
6	28	89.01%	93.87%
7	8	91.39%	94.91%
8	**1**	**95.56%**	**96.30%**

Table 5.3: Classification performance of our CSP classifier on the Weizmann human action dataset, using different numbers of training subjects. Combinations is the evaluated number of subsets of subjects.

Clearly, performance decreases with a decreasing amount of training data. But, even when only a few subjects are used for training, the results are reasonable. We expect that the variation in the direction of movement of the jump, run, side, skip and walk sequences will have a significant impact on the results, especially for the evaluations with very few training subjects. Even though we do not model the movement in the image, changing the direction of movement results in image observations to be mirrored. Of course, this results in very different silhouette descriptors. We will look at this issue further in the next section. Nevertheless, our approach can cope with these variations to some extent.

5.4.3 Results on robustness sequences

The Weizmann human action dataset contains additional robustness sequences that can be used to investigate how well an approach performs with less than perfect data. There are two types of sets, each of which contain 10 additional walking sequences. In the deformation sequences, different variations of walking are viewed from the side (see Figure 5.3 (top row)). These sequences include walking with objects (bag, briefcase, dog), different walking styles (kneesup, limp, moonwalk), different clothing styles (skirt) and occlusion settings (nofeet, pole). It is arguable whether the different styles should be classified as walking since they show many similarities with the skip action. In this experiment, we maintained to proposed labeling.

The viewpoint sequences show one walking subject, viewed from 0° (side view) to 81° (near-front view), in increments of 9°. Figure 5.3 (bottom row) shows example frames.

Figure 5.3: Example frames from the Weizmann robustness sequences. (top) Deformations, images and silhouettes for bag, briefcase, dog, kneesup, limp, moonwalk, nofeet, normwalk, pole and skirt. (bottom) Different viewpoints, images and silhouettes, 0° - 81° in increments of 9°.

Our experimental setup was similar to the one used earlier but we used the training data of all 9 subjects. We performed the experiments on the deformation and viewpoint sequences separately. We used the HOSG descriptors described in Section 5.2. Our results are averaged over the 10 sequences of each set. For the deformation sequences, we obtained 80.00% correct estimates. The incorrectly classified sequences were moonwalk and pole, both of which were classified as running. For

the viewpoints sequences, 80.00% were also classified correctly. The most challenging trials corresponding to viewpoints 72° and 81° were both classified as pjump.

When we reduced the number of subjects in our training set, we obtained lower results. Specifically, for 5 subjects, we scored 79.60% correctly on the deformations, and 70.16% on the viewpoints. For training on a single subject, these numbers decreased to 58.89% and 48.89%, respectively. These percentages are averages of all combinations of training subjects. For the condition where we test only on a single subject, we can evaluate the influence on walking direction on the performance, as the sequences in both the deformations and viewpoints sets show walking from left to right. When the training subject is walking in the same direction as the test subject, the scores are respectively 80.00% and 74.00% on the deformations and viewpoints sets. For the opposite direction, these numbers are significantly lower at 32.50% and 17.50%, respectively. Here, we did not look at the direction of related classes such as run and skip but it shows that it is important to take the direction of movement into account during training. Alternatively, different directions can be treated as different action classes.

5.4.4 Results on subsequences

So far, we have used the entire sequence for classification. We assumed that temporal segmentation of the action was performed previously. This raises the question as to how well our approach would perform when such accurate segmentation is not available. Since the Weizmann human action dataset contains only sequences with a single action, we focus on subsequences instead. We repeated our main LOOCV experiment but varied the length of the test sequences. The training phase was exactly the same, so we used the entire sequences. For the testing, we used a sliding window with a length in the range $[1, 25]$. The minimum sequence length was 28 frames. We slid the window through the sequence with steps of 1 frame. Average performance results over all sequences for different subsequence lengths are given in Figure 5.4(left). It is clear that increasing subsequence length results in an increased performance. This can be explained by the additional information that is available as the sequence becomes longer.

We expect that the relative progress within the sequence influences these results. Most action performances start and end in a resting pose. Also, for moving actions (e.g. walking and running), the start and end of the sequence take place partly outside the viewing window. Therefore, we looked at the classification performance at different relative times within the sequence. We calculated the relative start of the subsequence as the starting frame divided by the sequence length. To compare sequences of different lengths, we binned these values into a 10-dimensional histogram. Each cell contains the average classification performance. Figure 5.4(right) shows these results, averaged over all sequence lengths. It immediately becomes clear that performance is indeed lower at the start and at the end of the sequence.

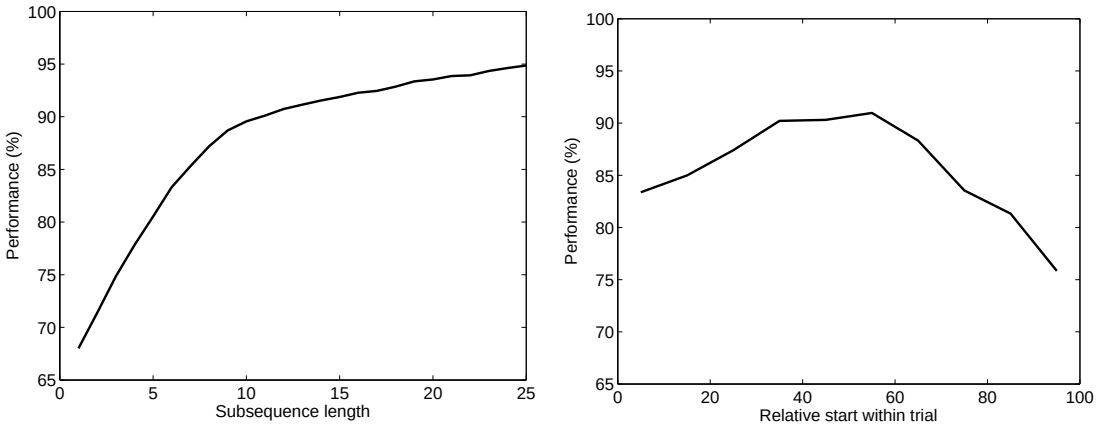


Figure 5.4: Classification performance for different subsequence lengths (left) and at different relative times (percentages given) within the sequence (right).

5.5 Discussion

In this section, we compare our results with those reported previously in literature. In Section 5.5.1, we present an in-depth comparison with recent exemplar-based holistic work. In Section 5.5.2, we compare our approach with other results on the Weizmann human action dataset. We discuss our approach with its strengths and its limitations in Section 5.5.3.

5.5.1 Comparison with exemplar-based holistic work

In many cases, humans can recognize human actions from only a single prototypical pose. Motivated by this observation, we explored the use of such key poses. Recently, Weinland and Boyer [385] presented an approach where they described sequences as a vector of minimum distances to selected exemplars. There are several approaches to select these exemplars. Unsupervised clustering algorithms such as k-means and expectation-maximization [65] are likely to select as exemplars those frames that are common among all classes. As such, they are not discriminative. Alternatively, the exemplar selection problem can be regarded as a feature subset selection problem, where each frame is a feature. There are three types of supervised approach [31; 122]. Filters select subsets as a preprocessing step, without taking into account the induction algorithm (classifier). Wilson and Martinez [389] present an overview of filter approaches. In contrast, wrapper approaches [177] explicitly use the induction algorithm into the subset selection scheme. A third approach is that of embedding methods, which perform subset selection within the training process. Usually, these methods are specifically designed for a given classifier and we do not consider them here.

In this section, we describe our implementation of the approach of Weinland and Boyer [385], using either k-medoids (k-means where cluster centra correspond to the closest exemplar) or the wrapper approach to select exemplars. We used a Bayes classifier where each class is described as a multivariate Gaussian. Given the conceptual

advantages of the wrapper approach over the unsupervised k-medoids, we expect to achieve higher accuracies for a smaller number of exemplars.

We used a LOOCV approach where each fold corresponds to one of the 9 test subjects. Our settings corresponded to those in [385], which we summarize here for completeness. Specifically, we used a forward selection scheme. We started with an empty set of exemplars $E = \emptyset$, and a full set of candidates $C = \{c_i | 1 \leq i \leq n\}$ with n the total number of candidates. We sampled $n = 300$ frames from the training sequences. At each iteration, an exemplar from the candidate set is moved to the exemplar set. This is the exemplar that results in the largest performance increase on the validation set. To make sure exemplar selection was not biased on a single subject, we used cross-validation within this exemplar-selection step. Since a perfect performance score on the validation set is easily obtained, we temporarily and randomly removed exemplars until the validation score was below 100%. In the validation step, we used the Bayes classifier where each class was described as a multivariate Gaussian. To avoid singularity problems in the inversion, we used an axis-aligned covariance matrix, in which all off-diagonal covariance elements are zero. We used Mahalanobis distance D to determine the distance of each unseen trial to all classes: $D = (x - \mu)^T \Sigma^{-1} (x - \mu)$, where x is the k-dimensional vector of minimum distances to the k selected exemplars, and μ and Σ are the mean and covariance of the given class, respectively.

When multiple frames resulted in the highest performance increase on the validation set, we randomly selected one of them. Also, the selection of the candidate set was random. Therefore, we present our results as averages over 3 repetitions. We used the HOSG-4×4 descriptors as our image representation, and performed our experiments with all 10 action classes.

Figure 5.5: Exemplars selected using k-medoids (top), and the wrapper approach (bottom), both with the number of exemplars $k = 10$. Test subject is *Daria*.

The results for different numbers of k are presented in Figure 5.6, with either k-medoids or the wrapper approach for exemplar selection. The graphs show that for the wrapper approach, performance increases more rapidly. This can be understood by the discriminative selection in the wrapper approach. Also, the performance with the wrapper approach is slightly higher. For one repetition, the exemplars for $k = 10$ are shown in Figure 5.5. The exemplars that are selected in the wrapper approach correspond more clearly to different classes, whereas k-medoids selects more exemplars that are common among classes, or are noisy. Confusion matrices for $k = 50$

are presented in Table 5.4. It is clear that run and skip are often guessed, which can be explained by the large within-variance. Remarkably, the two wave actions are both often classified as bend. This is probably due to scaling the bounding box to a fixed ratio. Especially for the bend and wave actions, the height of the human figure changes quickly. Notice the perfect recognition for the walk action when the wrapper approach is used. This shows the discriminative effect of the selected exemplars (see exemplar 1 and 7 in the bottom row of Figure 5.5).

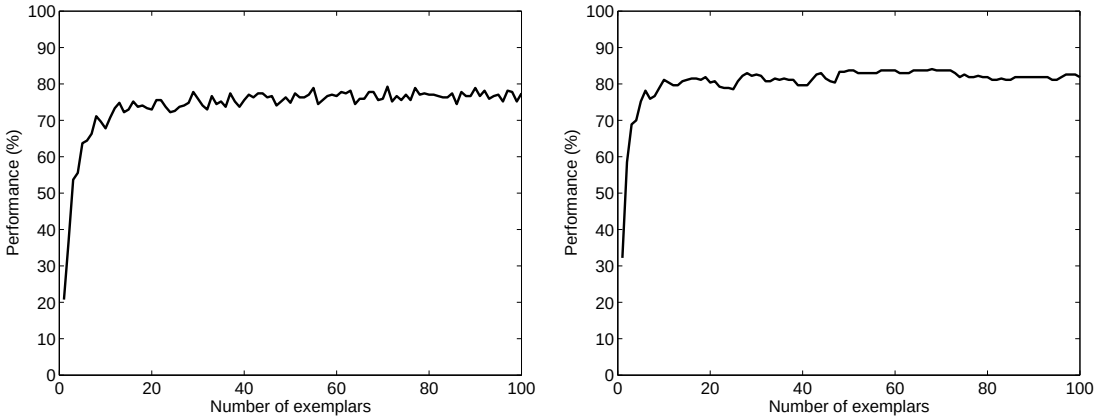


Figure 5.6: Classification performance for different numbers of exemplars k for k-medoids (left) and the wrapper approach (right).

	Guessed									
Actual	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	23				3		1			
jack		25			2					
jump	2		15		6		4			
pjump				24	3					
run					24		3			
side					3	24				
skip					18		9			
walk					10		1	16		
wave1	6				1				20	
wave2	4				1					22

	Guessed									
Actual	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	24						3			
jack		27								
jump			18				9			
pjump				25	2					
run					24		3			
side					3	24				
skip					14		13			
walk								27		
wave1	6								21	
wave2	3					1				23

Table 5.4: Confusion matrices for exemplar-based experiment, with $k = 50$ exemplars. Exemplars are selected using k-medoids (left, performance 74.81%) and the wrapper approach (right, performance 83.70%). The numbers are accumulated for all test subjects in 3 repetitions.

Both the exemplar-based approach and our CSP classifiers are discriminative but their strengths are different. The results of the exemplar-based approach are easily interpretable and arbitrary distance measures between frames and exemplars can be used. For example, Weinland and Boyer [385] use Chamfer distance and achieve 100% accuracy when at least 120 exemplars are used. The CSP classifiers are limited in that they require a vector representation. However, the CSP classifiers can be trained very efficiently and have been shown to yield good results even for small subsets or when limited training data is available. In a direct comparison using the

HOSG descriptors our CSP classifier outperforms the exemplar-based approach with over 10%. Differences between our results on the wrapper approach and those reported in [385] can be explained by the different image representation and matching. The Chamfer matching is more robust at the cost of being more computationally expensive.

5.5.2 Comparison with other related research

There have been several other reports of results on the Weizmann human action dataset. We review these and point out differences with our work. Such comparisons reveal the relative advantages of one method over the other. We selected works that are representative of a class of approaches.

Niebles and Fei-Fei [238] achieved a 72.80% score over 9 actions. Spatial and spatio-temporal interest points were sampled, and combined into a constellation. Action classification was performed by taking a majority vote over all individually classified frames. No background segmentation or localization was needed. This makes their approach more robust than ours. Recent work by Thurau [345] used HOG-descriptors for both detection and action classification. No background segmentation was used, but centered and aligned training data was needed. For classification, n-grams of action snippets were used. With all 10 actions and 90 bi-grams, performance was 86.66%.

In theory, the work of İkizler and Duygulu [144] did not require background segmentation but localization was assumed. A large number of rotated rectangular patches were extracted, and divided over a 3×3 grid, forming a histogram of oriented rectangles. A number of settings and classification methods was evaluated on the dataset without the skip action. All actions were classified correctly when using Dynamic Time Warping. This requires the temporal alignment of each unseen sequence to all sequences in the training set, which is computationally expensive. Using one histogram per sequence, 96.30% was scored. Again, this requires comparison to all training sequences. For comparison, we calculated the performance of our descriptor using a length-normalized histogram over the entire sequence and 1-nearest neighbor using Euclidian distance, and with the skip action left out. This resulted in 96.30% performance, a similar score.

Other works require background subtraction and used the masks that are provided with the dataset. Wang and Suter [376] scored 97.78% over all 10 actions. Raw silhouette values were used, and long-term dependencies between observations were modeled in their FCRF. When small blocks of pixels were regarded, thus effectively reducing the resolution, performance decreased. For 4×4 blocks and 8×8 blocks, scores were obtained of 92.22% and 77.78%, with descriptor sizes 192 and 48, respectively. Kernel PCA was used to reduce the dimensionality, but the dimension of the projected space was not reported. In contrast, we started with a 128-dimensional silhouette descriptor, and performed the classification using only 10 components. Moreover, our training requirements are much lower. On the other hand, FCRFs are able to model complex temporal dynamics.

There are several reports of schemes where subsequences are classified. For example, Blank *et al.* [30] used subsequences of 10 frames, and obtained a performance

of 99.64%. They used local features, extracted from a space-time volume that was constructed by concatenating silhouettes over time. Schindler and Van Gool [306] used local shape and optical flow, and evaluated their approach using subsequences between one and 10 frames. Their performance of 93.5% for a single frame increased to 99.60% when 10 frames were used. In contrast to these works, we used a holistic representation and no motion information. Such a representation can be obtained much faster. The downside is our lower performance of 89.56% using 10-frame subsequences.

5.5.3 Conclusion

We have shown that the application of common spatial patterns to increase the margin between pairs of classes, also increases classification performance. Confusions that remain are between related classes such as walking and running. These results are competitive, and we have shown that we can even obtain reasonable results with only a few training subjects. Moreover, training and evaluation complexity are low.

Oriented silhouette boundaries were encoded as a histogram of orientated gradients (HOSG) within cells of a grid. Such a holistic representation can be calculated fast but generally cannot cope with more severe occlusions. We conducted additional experiments with walking sequences with variations. Here, it was shown that the sequence with a pole-like occlusion was not classified correctly. Another drawback is the dependence on an accurately determined region of interest. To assess the performance of our method on more realistic scenes reliably, our work should be combined with an automatic human detection preprocessing step, such as in [414].

Simple pairwise discriminative functions were used, where each class was represented by an average vector of all training sequences of the class. Such an approach is simple, yet does not model intra-class variance. Such a prototype is likely to be an average of multiple modes, especially when there are large differences within the class, such as different directions of movement. To overcome this issue, multiple classes for a single action could be introduced, depending on the direction of movement. Moreover, the temporal aspect in our action prototypes is, to a great extent, ignored. Performance could be increased by including temporal characteristics.

We have evaluated our work on entire sequences and subsequences. We did not explicitly address the temporal segmentation. Also, current datasets for human action recognition do not contain an ‘other’ class. Instead of selecting the class with the highest voting mass, this would also require an approach to decide whether the chosen class is really observed. Generally, this is a harder problem since there is more variation in the ‘other’ class and the prior probabilities for the classes can vary significantly.

6

Human action recognition from recovered poses

In the previous chapter, we presented our approach to classify human movement from image sequences. While we achieved good distinction for side views, the main limitation of this approach is the dependency on the viewpoint. We showed that recognition of walking motion was unaffected for moderate changes in viewpoint, but larger differences in viewing direction between training and test sequences could not be accommodated. Moreover, we expect that for actions with more variation (such as throwing a ball), it might prove to be more difficult to recognize the action from different viewpoints. In Chapter 3, we introduced our example-based approach to human pose recovery. We demonstrated that we could recover poses with reasonable accuracy from image descriptors. Our pose descriptor consisted of the 3D locations of 20 key joints in the human body relative to the pelvis. We believe that such a representation is sufficiently rich to allow for recognition of human actions. Therefore, in this chapter, we combine our pose recovery approach with our common spatial patterns classifier to recognize human actions from recovered poses instead of from image descriptors.

One great advantage of using poses is that they can be made invariant to rotations. As such, we can train and test on sequences recorded from different viewpoints. We still require that a certain pose from a certain viewpoint can be recovered but we do not require a performance of the action from a specific viewpoint. Moreover, we show that we can learn action models from motion capture data. We can use these models to test on sequences of recovered poses. Ideally, one could record motion capture data once and learn the action models from this data. In addition, character generation software such as Poser, Maya or 3D Studio Max could be used to generate images from a variety of viewpoints. The corresponding image descriptors together with the associated poses could form the example database to be used for pose recovery. Such an approach would enable simultaneous recovery of human poses and recognition of the action from arbitrary viewpoints with low training requirements. Instead of synthetically generated images and in line with Chapter 3, we use real images from the HumanEva dataset in this chapter.

Another advantage of our combined approach is the reduced sensitivity to partial

occlusions. In Section 3.3, we demonstrated that we could recover poses under partial occlusion. Since we use these recovered poses instead of image descriptors, we can recognize actions when these occlusions are present. This is not possible when using the image descriptors directly, such as in the previous chapter. The application of PCA will include the dimensions, corresponding to orientation bins of the occluded cells, into multiple (usually all) transformed components. This effect is also present for the selection of the CSP components.

We observed in the previous chapter that our discriminative CSP classifiers can be trained with very few training sequences. In this chapter, we use the HumanEva dataset, which contains a significantly lower number of sequences than the Weizmann human action dataset. Consequently, some actions that we want to recognize have only a single or small number of training sequences available. Nevertheless, we demonstrate that we can obtain reasonable results for the classification of human actions.

As both the pose recovery approach and the CSP classifier have low computation requirements, our combined approach is still fast. With an optimized implementation, our approach can be made suitable for real-time simultaneous recovery of human poses and recognition of human action. For online human action recognition, there will always be a short delay due to the fact that we use short sequences of movement over time for classification.

Human pose descriptors have been used for human action recognition in several different classification approaches. For example, Ali *et al*. [10] use the approximate 2D locations of the head, hands and feet. These can be recovered relatively easily but do not allow for viewpoint-invariant action recognition.

Also, the use of 3D pose descriptors to train and test action models is not novel. However, several other works do not regard the recovery of the poses, but assume that accurate pose descriptors are available [262; 285; 313; 314; 405]. Recent work by Han *et al*. [123], is somewhat related to our approach as they also take a discriminative approach. They first project the pose representation down using a Hierarchical Gaussian Process Latent Variable Model (HGPLVM). Next, they learn motion patterns for each limb individually and use a CRF to predict the motion in the manifold subspace. Finally, an SVM is used to classify the motion pattern. Their approach is limited as they only consider movements with low intra-class variation. Also, temporal segmentation is assumed and it remains unclear how the approach would perform with different rates of movement. Also related is the work by Lv and Nevatia [206], where the motion of each limb is modeled with an HMM and AdaBoost is used to discriminate between actions. While some of the above approaches are invariant to body dimensions and speed of performance, it remains an open question how these methods would perform with less than perfect pose estimates, for example, due to pose recovery inaccuracies.

The combination of pose recovery and human action recognition has been addressed in a number of works. The limb tracker in Ramanan *et al*. [283] can recover 2D joint locations without assuming a particular motion model. Their work is based on the pictorial structures idea of [89]. In later work, 2D joint tracks are lifted to 3D using stored motion capture fragments [282]. These fragments have been annotated

in terms of action labels. Action recognition is based on the frequency of occurrence of these annotations in a recovered motion sequence, and an HMM is further used to smooth these estimations over time. İkizler and Forsyth [145] use a slightly different approach where lifting of the 2D tracks to 3D is performed for each limb separately. Recent work by Ferrari *et al*. [93] uses a progressive reduction of the search space to find upper-body poses. They first apply a pose-independent upper-body detector that is similar to the HOG-based person detection work by Dalal and Triggs [58]. After this reduction, they use the pictorial structures concept to iteratively determine the body pose [281]. The segmentation of the body pose is regarded as a richer descriptor and subsequently used for action recognition. These works all rely on the bottom-up detection of limbs. Such an approach has the advantage that no appearance model is required. Also, no background segmentation step needs to be performed. The drawback is in the computational complexity. Both the recovery of the 2D poses, and the lifting to 3D are computationally expensive processes and it will be a challenge to perform these tasks in real time. Work by Peursum *et al*. [268] combines pose recovery and action recognition using a variant of the hierarchical HMM. While their approach can deal with many real-world conditions such as partial occlusions and generalization between subjects, their approach is rather slow due to the high parameter space.

In contrast, our work has a low computational complexity. Apart from a small delay caused by the fact that we classify motion over a short span of time, our approach can be made to operate in real time. We achieve this by combining our fast pose recovery approach with fast classification of human actions. To the best of our knowledge, there is no work that attempts the online, simultaneous recovery of human poses and recognition of actions. In addition, we show that our approach can deal with partial occlusions when these are predicted. Moreover, our experiments demonstrate that we can deal with considerable variation of movement even when we have only a small number of training sequences available. Finally, we show that we can temporally segment sequences of movement. A limitation of our approach is that we assume detection of the human figure obtained through background segmentation. Also, our example database requires a reasonable number of examples that span the convex hull of those poses and viewpoints that we expect to recover.

In Section 6.1, we discuss the adaptations that we need to make in order to be able to use our human action recognition approach with pose descriptors. Specifically, we explain how we normalize the pose descriptors for rotation and differences in human body dimensions. We present our main action recognition experiment in Section 6.2, and describe additional experiments and their results in Section 6.3. Finally, we discuss our approach with its strengths and limitations in Section 6.4.

6.1 Adaptations to the action recognition approach

When we replace the image descriptors with pose descriptors, we do not have to adapt the common spatial pattern approach that we described in Section 5.1. Specifically, we determine and evaluate the discriminative functions in exactly the same way. The type of input is all that is changed. In the classification of human movement, we are not interested in the direction in which a certain action is performed. This implies

that we do not distinguish between, for example, walking in a circle and walking in a straight line. While for some applications, this information is valuable, training complexity is much lower when the direction is ignored. Optionally, the heading direction can be used at a later stage. Here, we do not employ such a step. Apart from variation in the direction of movement, there are also variations in body size between persons. Given that our CSP classifier uses the variance in joint locations, differences in body dimensions affect the CSP transform. To reduce this influence, we apply uniform normalization on all joint distances.

We discuss the rotation normalization in Section 6.1.1, and the rotation for different body heights in Section 6.1.2.

6.1.1 Rotation normalization

We use rotation-invariant pose descriptors in order to be able to train and test action models regardless of the direction of the movement. To this end, we normalize the rotation of the poses. We only consider rotations around a vertical axis, as a rotation around this axis does not affect the relative direction of gravity. Consequently, two poses that are equal, apart from a rotation around a vertical axis will be performed with the same muscular activity. This is not the case for any other axis. For example, we consider making jumping jacks (or star jumps) and making snow angels (the same movement but performed horizontally in the snow, usually by children) as two different activities whereas the direction in which either is performed, does not affect the labeling.

To normalize poses for this rotation, we need to determine the heading direction of the movement, or rather the direction in which the body is facing. Since this direction is not explicitly defined in the pose description, we need to define how we determine it. Ideally, we would use two joints that are approximately in the same plane, parallel to the ground plane. In this case, we can easily determine the angle between a line in the ground plane, and a the projection of a line between those two joints. Given the markers that are used in the HumanEva dataset, there are two options: the two shoulder joints (*upperRArmProximal* and *upperLArmProximal*) or the two hip joints (*upperLLegProximal* and *lowerLLegProximal*). We found that the shoulder joints showed considerable variation when performing the box, gesture and throw movements. Also, it appeared that the recovery errors for the shoulder joints were, on average, higher than those for the hip joints. Therefore, we use the hip joints to determine the heading rotation. To this end, we calculate the directed, inverse tangent of the positive x-axis and the line through the two hips joints projected onto the ground plane. Note that using the positive x-axis is an arbitrary choice and we could have used any line in the ground plane without affecting the results. Figure 6.1 shows several frames with recovered poses overlaid together with the normalized poses which are all facing the camera.

6.1.2 Body height normalization

There is considerable variation in the body height of the different subjects in the HumanEva dataset that are available for training. When using joint locations as action

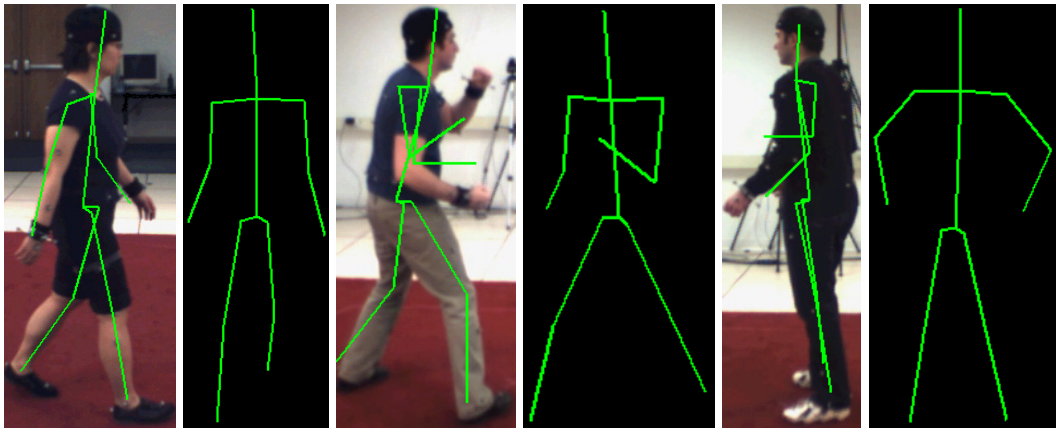


Figure 6.1: Recovered poses of frames 100 of different actions, evaluated with a single camera using T1. (**a**,**b**) walking, camera C1, subject 1 (**c**,**d**) box, camera C2, subject 2 (**e**,**f**) throw, camera C3, subject 3. For the directed poses in (**a**,**c**,**e**), the 3D pelvis location was set manually, as we only estimate relative joint locations. The rotation-normalized poses in (**b**,**d**,**f**) are scaled differently.

features, variation in body dimensions have an effect on the CSP components. CSP partly focusses on this variance between subjects instead of the variance within the movement itself. To solve this issue to some extent we normalize all joint distances. In practice, normalization of each segment individually would allow for optimal comparison of movements performed by different subjects. However, such a normalization requires robust determination of segment lengths from data, which is difficult to obtain due to inaccurate pose recovery. Instead, we use uniform scaling of all segment lengths. Since the height of a subject is not known, we have to use the recovered pose itself to obtain an estimation. After recovering the pose, the height of a person could be approximated by summing the lengths of lower and upper legs, the spine and the head. However, since these joint locations that are used to calculate these lengths are obtained through weighted interpolation, the locations of the feet can significantly differ from the real pose. The fact that the joint locations that are used within HumanEva do not exactly coincide with the physical joint locations adds to this problem. A bent leg could therefore have a different estimated length than a stretched leg. These variations in estimated leg length have an effect on the overall estimation of the subject's height.

We quantitatively analyzed the motion capture data of the HumanEva Walking training sequences. We calculated the subjects' length by summing the lengths of the left lower and upper legs, the spine and the head. The lengths of subjects 1, 2 and 3 are 153.19 *cm*, 162.83 *cm* and 170.20 *cm*, respectively. We have also calculated the lengths from poses that were recovered using the method described in Section 3.2, with the HOG-F-5×6 descriptors and using monocular example set T1. Again, we used only the Walking sequences. The average lengths of subjects 1, 2 and 3 are 153.18 *cm*, 162.82 *cm* and 169.46 *cm*. These values are very similar to those obtained from the motion capture data. While these numbers led us to believe that we could

robustly use the approximated height of the subject, there is significant variation in this height between frames. For example, the standard deviation of the estimated height of subject 1 is 16.33 *mm* (16.12 *mm* when using mocap).

If we look at the length of the spine alone, this variance is much lower. For subject 1, we obtain 1.92 *mm* when using recovered poses, and 1.60 *mm* when using motion capture data. The average subject's height is approximately four times the length of the spine, so an equal deviation would have more impact when using the spine for normalization. However, the much lower variation in estimated spine length motivates our decision to use the spine length for normalization. This choice has another advantage as we do not make use of the length of the head. As we mentioned before in Section 3.2.3.3, some of the training sequences of subject 3 contain erroneous values for the *headProximal* marker, which is located on top of the head. While examples with these values still influence the estimation of this marker, they do not influence the normalization.

6.2 Experiment results

We evaluate the performance of our combined approach on the HumanEva-I dataset. The use of this dataset is suitable as it allows us to look closer at the influence of inaccuracies in the recovery of poses on the classification results. Moreover, the actions in the HumanEva dataset have been performed in a less controlled manner. Compared to the Weizmann human action dataset that we used in the previous chapter, there is much more variation in the performance of the different actions. There is considerable variation in action execution for different subjects. This is especially true for the non-cyclic actions such as punching and catching a ball. Figure 6.2 shows several frames of the right punch action where the arm is maximally extended. Apart from these different poses, there is much variation in the movement as well. This makes the dataset more realistic than, for example, the Weizmann human action dataset.

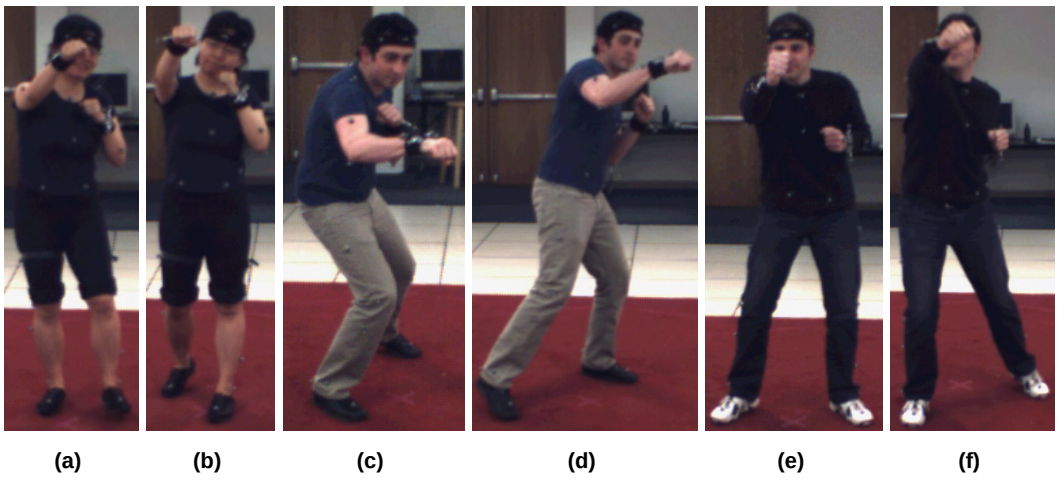


Figure 6.2: Example frames of the punch right action, performed by (**a**-**b**) subject 1, (**c**-**d**) subject 2, (**e**-**f**) subject 3. Frames (**a**,**c**,**e**) are from the training set, frames (**b**,**d**,**f**) are from the test set. All frames are viewed from camera C1.

Each sequence in the HumanEva dataset contains one activity that can consist of several different actions. For example, a Throw/Catch sequence can consist of several throws and catches and some additional segments where the subject is in rest or walking. Given that these actions are performed in sequence, there is a transition from one action to the other. While it is difficult to determine where one action stops and the other starts, such a transition is typical for realistic human movement. In our experiments we have temporally segmented the actions but we have not explicitly labeled these transitions. As a result, these transitions are part of the performance of the action, which also adds to the variation within an action class. Additionally, we present an experiment in Section 6.3.3 where we investigate how well we can use our approach to perform temporal segmentation of the human motion.

Another important motivation for the use of the HumanEva dataset is the fact that there is considerable variation in the direction of performance. The Walking and Jog activities are performed in a circle but there is also moderate variation for many other actions (see also Figure 6.2). We only present our results using camera C1, but our approach would not require retraining if we were to use any of the other cameras.

To the best of our knowledge, we are the first to perform more specific action recognition on the HumanEva dataset. Ning *et al*. [245] use the HumanEva set but only distinguish between walking, jogging, boxing and gesturing. In contrast, we distinguish between actions with more subtle differences. The INRIA XMAS multi-view dataset (see Section 4.1.4.3) contains multiple viewpoints, and each sequence consists of a number of actions performed in sequence. However, no pose information is present. Also, the different actions are performed in a fixed order which makes the transitions less realistic. Since the overlap in actions of the different human action datasets is minimal, we have only used the HumanEva dataset in our experiments.

We describe the construction of the training and test sets from the HumanEva sequences in Section 6.2.1. The setup of our experiment, where we train and test on poses that have been recovered using our example-based approach, is explained in Section 6.2.2. In Section 6.2.3, we summarize and discuss our results. Additional experiments are presented in Section 6.3.

6.2.1 HumanEva action dataset

The sequences in the HumanEva-I dataset are labeled by one activity (Box, Gesture, Jog, Throw/Catch, Walking, Combo) but contain a variety of actions. Therefore we manually labeled the training and test sequences with action labels. A successive number of frames with the same action label is termed an action segment and can contain multiple iterations of the action. Table 6.1 summarizes the number of action segments and the total number of frames for both the training sequences and the test sequences. The frame rate of the HumanEva dataset is 60 frames per second. In line with Chapter 3, we only use the HumanEva-I sequences of subjects 1-3. There are only a few segments of the walking and jog action which are all relatively long. This is because these actions are performed repeatedly without interruption of another action. In contrast, the average number of frames of the throw and box actions is much lower due to the fact that different actions are performed in rapid succession.

The labeling of the action segments is arbitrary. First, the label set is a compromise

between specificity and generalization. For example, there is much variation in the boxing and gesture actions, such as in the direction of the punch or in the height of the wave gesture. Yet, we only distinguish between the hand that makes the movements, and between two coarse movement classes. Second, the temporal segmentation could have been performed differently as it is often not clear where in time an action starts and ends. As noted before, the sequences contain transitions from one action to the next. We have not explicitly labeled these transitions but rather, each action segment starts with approximately half of the transition from the previous action, and ends with the first half of the transition to the next action.

Table 6.1 shows that some actions are only performed by one or two subjects. Some of these have only been performed once, such as the clap, wave both and get up actions. The balancing and jump actions in the Combo sequences are only available in the test set. For other actions, there are only very short segments available. In our experiment we evaluate the human action recognition performance for several different sub-sequence lengths. In some cases we do not have any sufficiently long training or test segments available. Therefore we use different subsets of the action classes depending on the sub-sequence length setting. Our only criterion for including an action in the subset is that there is at least one training segment and one test segment of sufficient length available for the action.

For both the training and the test segments, we recovered the pose for each frame. To this end we used our example-based pose recovery approach that is described in detail in Chapter 3. We used the same settings, with the HOG-*F*-5×6 descriptors, monocular example set T1 and all frames viewed from camera C1. When a training frame had valid motion capture data associated, the recovered pose was exactly the same as the motion capture data. This is the result of using the weighted normalization in the recovery of the poses. For frames where the motion capture data was invalid, a weighted interpolation of valid poses from the example database was used. In contrast to only using the valid motion capture frames, the use of recovered poses allowed us to use all frames in the training sequences. Were we only to consider frames with valid motion capture, the available number of sub-sequences of sufficient length would be drastically reduced.

We performed the rotation and body length normalization as described in the previous section for the recovered pose of each frame. The pelvis (*torsoDistal*) joint is always located at the origin of a local coordinate space. Therefore, we removed this joint from the pose description, which reduced the dimensionality to 57D.

6.2.2 Experiment setup

In this section we discuss the setup of our main experiment. We used separate training and test sets, in contrast to the previous chapter where a leave-one-out approach was adopted. Also, instead of image descriptors, we used the recovered poses of the HumanEva-I action as explained in the previous section.

Even after the removal of the *torsoDistal* joint from the pose description, some of the dimensions show low variance, such as the hip and neck joints. To avoid singularity problems in calculating the inverse covariance matrices, we selected the first 30 out of 57 components after performing principal component analysis. These com-

	Training			Test		
Action	S1	S2	S3	S1	S2	S3
Rest	5 (453)	7 (624)	8 (373)	5 (370)	13 (780)	10 (461)
Walking	1 (1203)	2 (974)	1 (939)	3 (2162)	4 (2078)	3 (1515)
Jog	1 (740)	1 (795)	1 (842)	2 (1603)	2 (1328)	2 (1390)
Punch r.	5 (225)	7 (352)	9 (448)	3 (163)	7 (348)	5 (210)
Punch l.	4 (200)	6 (258)	9 (328)	3 (120)	6 (335)	5 (187)
Uppercut r.	2 (89)	1 (39)	3 (195)	3 (146)	2 (134)	2 (133)
Uppercut l.	2 (136)	1 (45)	1 (36)	3 (177)	0 (0)	0 (0)
Wave r.	4 (449)	2 (100)	5 (511)	5 (516)	4 (252)	2 (224)
Wave l.	0 (0)	1 (80)	0 (0)	0 (0)	2 (117)	0 (0)
Beckon r.	3 (352)	4 (301)	4 (516)	5 (554)	4 (346)	3 (329)
Beckon l.	0 (0)	2 (128)	0 (0)	0 (0)	1 (93)	0 (0)
Throw low r.	2 (139)	1 (108)	3 (240)	1 (104)	3 (322)	3 (240)
Throw side r.	0 (0)	1 (104)	0 (0)	1 (83)	0 (0)	0 (0)
Throw high r.	2 (180)	1 (120)	2 (172)	2 (175)	2 (193)	2 (169)
Catch	4 (243)	3 (246)	5 (317)	4 (219)	5 (358)	6 (353)
Clap	1 (69)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Wave both	0 (0)	1 (89)	0 (0)	0 (0)	0 (0)	0 (0)
Get up	0 (0)	1 (71)	0 (0)	0 (0)	0 (0)	0 (0)
Balance foot r.	0 (0)	0 (0)	0 (0)	1 (122)	1 (138)	1 (186)
Balance foot l.	0 (0)	0 (0)	0 (0)	2 (203)	1 (137)	1 (192)
Jump foot r.	0 (0)	0 (0)	0 (0)	1 (198)	1 (98)	1 (66)
Jump foot l.	0 (0)	0 (0)	0 (0)	1 (307)	1 (95)	1 (118)
Rest hands high	0 (0)	0 (0)	0 (0)	0 (0)	3 (172)	0 (0)

Table 6.1: Number of segments for different actions and subjects in the training and test set of the HumanEva-I dataset. Numbers between brackets are the total numbers of available frames. Abbreviation r. stands for right, l. stands for left.

ponents explained approximately 98% of the variance. For all experiments described in this chapter, we used 6 ($k = 3$) CSP components to describe the action prototype vectors and the test sequences. We experimented with other values for k but found no improvement for $k > 3$. In line with our experiments with image descriptors, this is 20% of the dimensionality after PCA reduction. Again, each discriminant function softly voted into the two classes using Equation 5.5. We selected the class that received the highest voting mass as the estimated action class.

There is a lot of variation in the duration of the performance of an action. A single repetition of a wave action lasts approximately half a second, whereas a throwing action takes on average three times longer. Also, many walking, jog, wave and beckon segments contain multiple repetitions of the action. If these were to be treated as a single segment, there would be much less data available for training and testing. Therefore, each segment was divided into sub-sequences of fixed length. Specifically, we used sub-sequence lengths of 30, 60, 90 and 120 frames. The frame rate of the HumanEva dataset is 60 frames per second, thus these lengths range from half a second to two seconds. To increase the number of training and test sub-sequences, two

adjacent sub-sequences overlapped with half of their length. As mentioned in the previous section, the sub-sequence length influences the number of sub-sequences that are available for training and testing. These numbers are summarized in Table 6.2. Given that we required at least one sub-sequence for training and one for testing, the total number of action classes was 15 for sub-sequence lengths of 30 and 60, 7 for lengths of 90 frames, and 5 when 120 subsequent frames were used.

Sub-sequence length	30 frames		60 frames		90 frames		120 frames	
Action	Train	Test	Train	Test	Train	Test	Train	Test
Rest	68	65	21	11	10	5	5	2
Walking	202	368	99	178	63	111	46	82
Jog	155	280	75	136	48	88	36	64
Punch r.	39	27	5	3	2	0	0	0
Punch l.	25	23	3	1	0	0	0	0
Uppercut r.	13	18	2	4	0	0	0	0
Uppercut l.	10	7	2	1	0	0	0	0
Wave r.	56	51	19	18	8	8	2	2
Wave l.	4	5	1	1	0	0	0	0
Beckon r.	60	65	21	22	10	11	3	3
Beckon l.	6	5	1	2	0	0	0	0
Throw low r.	24	34	7	11	2	4	0	0
Throw side r.	5	4	2	1	1	0	0	0
Throw high r.	25	27	9	9	3	3	1	0
Catch	36	38	9	8	0	0	0	0

Table 6.2: Number of available training and test sub-sequences for different sub-sequence lengths.

Training of the CSP classifier proceeded in the same manner as when using image descriptors. Specifically, for each pair of action classes the action prototypes were the means of the transformed vectors of all training sub-sequences. From the action prototype vectors for each pair of classes, we constructed the discriminative functions as in Equation 5.5.

6.2.3 Results

The human action recognition results of the CSP classifier on the HumanEva action sub-sequences are summarized in Table 6.3. For comparison, we also calculated the performance without the use of CSP both with the 30 and the 6 first PCA components. It is clear that the CSP classifier performs better than when the pose descriptors are used directly.

When comparing the different sub-sequence lengths, we observe an increase in performance for longer sub-sequences. Part of this increase can be attributed to the fewer number of classes for sub-sequence lengths of 90 and 120 frames. For an uninformed guess, the baseline is 14.29% and 20% for these lengths, respectively, instead of the 6.67% for the two shorter sub-sequence lengths. If we would know the *a priori* class probabilities, we could make an informed guess and always choose the class

with the highest probability. In this case, the baselines for sub-sequence lengths of 30, 60, 90 and 120 frames are 36.18%, 43.84%, 48.26% and 53.59%, respectively. However, we do not explicitly use information about *a priori* class probabilities.

Another important factor is the high number of walking and jog sub-sequences. In general, these were classified correctly more often then other classes. This is the case both when using CSP, and without CSP. For sub-sequence lengths of 30 frames, the total number of walking and jog sub-sequences is 64.35% of all sub-sequences. For a sub-sequence length of 120, this share is 95.42%. The relatively low score in the case with a sub-sequence length of 90 frames but without CSP can mainly be explained by the fact that many walking sub-sequences of subjects 1 and 2 were classified as rest action. The walking movements of subject 3 might be classified correctly due to the longer stride of this subject. Consequently, the extension of the arms and legs is more pronounced. A third factor that contributed to the higher performance is the fact that longer sub-sequences contain more information. The 30 frames, or half a second, often contain only part of the whole action. For example, a throw action can take two seconds. When an action is only observed for half a second, characteristic movement might not be taken into account. When longer sub-sequences are used, the probability that this characteristic movement is included in the sub-sequence is higher.

Sub-sequence length	30 frames	60 frames	90 frames	120 frames
Number of classes	15	15	7	5
CSP classifier	76.89%	87.20%	92.17%	93.46%
Without CSP (30D)	68.04%	71.50%	69.13%	81.70%
Without CSP (6D)	63.72%	65.46%	69.57%	71.24%

Table 6.3: Classification performance on the HumanEva-I human action dataset, for different sub-sequence lengths and corresponding action subsets.

We discuss the results for sub-sequence lengths of 60 frames in more detail. For this length, all 15 action classes were used. Confusion matrices of the experiment results when using the CSP classifier and without the CSP transform are presented in Tables 6.4 and 6.5, respectively. Our first observation is the high recall and precision of the walking and jog actions, especially for the CSP classifier. As these two actions were performed in a circle it is clear that our approach can deal with variations in viewpoint. In the latter table, we present results when using the first 6 principal components but without CSP transform. Compared to the CSP classifier, the main difference in performance was caused by the frequent misclassifications of the walking action. The vast majority of these confusions corresponds to sub-sequences performed by subject 1. Between the two settings, there are also differences in performance for the right wave and beckon actions. When no CSP transform was used, all beckon movements with the right hand were misclassified, mostly as waving with the right hand or catching a ball. With CSP, the performance for the beckon right action is 72.72%, and the confusions are all with the jog action. For the wave action, there are some confusions with the beckon action. When no CSP was used, there are some additional confusions with the throw low and catch actions.

There are clearly differences in both recall and precision for the different actions.

	Guessed																
Actual	Rest	Walking	Jog	Punch r.	Punch l.	Uppercut r.	Uppercut l.	Wave r.	Wave l.	Beckon r.	Beckon l.	Throw low r.	Throw side r.	Throw high r.	Catch	Recall	Precision
Rest	9	4	1	0	0	0	0	0	0	2	0	0	0	0	0	56.25%	33.33%
Walking	8	169	1	0	0	0	0	0	0	0	0	0	0	0	0	94.94%	96.02%
Jog	0	0	136	0	0	0	0	0	0	0	0	0	0	0	0	100.00%	91.89%
Punch r.	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	50.00%	66.67%
Punch l.	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0.00%	0.00%
Uppercut r.	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	50.00%	100.00%
Uppercut l.	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	100.00%	100.00%
Wave r.	0	0	0	0	0	0	0	13	0	5	0	0	0	0	0	72.22%	100.00%
Wave l.	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%	0.00%
Beckon r.	0	0	6	0	0	0	0	0	0	16	0	0	0	0	0	72.72%	66.67%
Beckon l.	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%	0.00%
Throw low r.	1	2	0	0	0	0	0	0	0	0	0	4	0	0	4	36.36%	80.00%
Throw side r.	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.00%	0.00%
Throw high r.	0	0	0	0	0	0	0	0	0	1	0	0	0	7	1	77.77%	87.50%
Catch	5	1	0	0	0	0	0	0	0	0	0	1	0	0	2	22.22%	28.57%

Table 6.4: Confusion matrix for the CSP classifier, trained and tested on recovered pose sub-sequences of length 60. Recall and precision values are given for each class.

The walking and jog actions have a high recall and precision, but some of the other actions were poorly recognized. This is especially true for the actions that were performed with the left hand, such as the punch, wave and beckon actions. For these actions, there were very few training sub-sequences available. Moreover, all these actions were performed by subject 2 who showed a lot of variation in his movements. This not only had an effect on the action classification itself but also on the prior recovery of the poses. We expect that this is generally the case. More variation in the performance of the movement will decrease recovery accuracy and consequently affect the action recognition. A similar effect is also present for the recovery of poses in the training set for which no valid motion capture data was available.

The large number of test sub-sequences for the walking and jog actions determined the performance reported in Table 6.3 to a great extent. Alternatively, we could present the average recall over all actions. Such a measure ignores the number of available test sub-sequences. This reduces the effect of the large number of walking and jog sub-sequences but at the same time puts a disproportional weight to action classes that have only a few test sub-sequences available. When using sub-sequences of 60 frames, the average recall over all classes is 48.83% for the CSP classifier, and 42.42% when no CSP transform was used. For comparison, for sub-sequences of 90 frames, and only 7 action classes, the average recall was 79.20% and 67.69%, respectively. This large difference can largely be explained by the fact that in the latter case actions with short segments were not taken into account. These were the actions that had only a few training sub-sequences available when using 60 frames.

	Guessed															
Actual	Rest	Walking	Jog	Punch r.	Punch l.	Uppercut r.	Uppercut l.	Wave r.	Wave l.	Beckon r.	Beckon l.	Throw low r.	Throw side r.	Throw high r.	Catch	Recall	Precision
Rest	8	0	2	0	0	1	0	0	0	2	3	0	0	0	0	50.00%	9.41%
Walking	74	104	0	0	0	0	0	0	0	0	0	0	0	0	0	58.43%	99.05%
Jog	0	1	135	0	0	0	0	0	0	0	0	0	0	0	0	99.26%	98.54%
Punch r.	0	0	0	1	0	0	2	0	0	0	1	0	0	0	0	25.00%	25.00%
Punch l.	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0.00%	0.00%
Uppercut r.	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	50.00%	25.00%
Uppercut l.	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	100.00%	25.00%
Wave r.	0	0	0	0	0	0	0	10	0	3	0	2	0	0	3	55.55%	47.62%
Wave l.	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	100.00%	25.00%
Beckon r.	0	0	0	1	0	1	0	10	0	0	0	0	0	0	10	0.00%	0.00%
Beckon l.	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0.00%	0.00%
Throw low r.	3	0	0	0	0	3	0	0	1	0	3	1	0	0	0	9.09%	25.00%
Throw side r.	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.00%	0.00%
Throw high r.	0	0	0	0	0	0	1	1	0	2	0	0	0	5	0	55.55%	71.43%
Catch	0	0	0	0	0	1	0	0	0	0	3	1	0	1	3	33.33%	18.75%

Table 6.5: Confusion matrix for the classifier without CSP transform, trained and tested on recovered pose sub-sequences of length 60. The first 6 principal components are used as pose descriptor. Recall and precision values are given for each class.

6.3 Additional experiments and results

In this section we present some additional experiments where we investigated the performance of our approach under different settings and conditions. In Section 6.3.1, we evaluate our approach on poses that are recovered from different HOG and HOSG descriptors. We present our experiment with simulated occlusion in Section 6.3.2. Finally, we demonstrate that our approach can be used for temporal segmentation of human action. A discussion of our approach and the results obtained appears in Section 6.4.

6.3.1 Results using different image representations

In the experiment described in Section 6.2, we used the HOG-*F*-5×6 descriptor to describe the image observation of the person. In this section, we evaluate the action recognition performance when using HOG and HOSG descriptors with different grid sizes. The settings we used are similar to those described in Sections 3.2.4.1 and 5.4.1. Specifically, we used the HOG-*R* and HOSG-*R* descriptors, where the ROIs have a fixed ratio between height and width, instead of the setting of the HOG-*F*-5×6 descriptor where the ROI exactly fits the silhouette. The use of different image descriptors results in different pose descriptors and, consequently, we expect that action recognition performance is affected.

The setup of this experiment was similar to the main experiment described in the previous section. Both training and testing were performed on the recovered poses obtained from different image descriptors. It should be noted that the training and test sequences could have been recovered with different image representations as

only the poses are used in constructing the classifiers. However, we have used the same representation. The number of training and test sub-sequences can be found in Table 6.2. We only evaluated sub-sequences of 60 frames in length. The corresponding number of action classes was therefore 15. The results of the experiment are summarized in Table 6.6.

	HOG-*R*			HOSG-*R*		
	3×3	4×4	5×6	3×3	4×4	5×6
CSP classifier	85.02%	86.47%	78.02%	64.73%	78.26%	80.92%
Without CSP (30D)	72.71%	73.19%	78.02%	74.40%	79.95%	82.13%
Without CSP (6D)	70.05%	65.94%	71.50%	64.25%	71.98%	71.26%

Table 6.6: Classification performance on the HumanEva-I human action dataset, for different HOG and HOSG grid sizes. The sub-sequence length was 60 frames, and 15 classes were used.

For the CSP classifier, the performance of the HOG descriptors was higher than that of the HOSG descriptors. To explain these differences, we look at the pose recovery accuracy for these image descriptors in Table 3.8. It appears that, for the CSP classifier, lower recovery errors resulted in higher action recognition rates. There is an especially high correlation between the recognition rate and the accuracy of the pose recovery for the Walking and Jog activities. The vast majority of the sub-sequences corresponded to these two action classes. Compared to the experiment with the HOG-*F*-5×6 descriptors, the action recognition rates in Table 6.6 are slightly lower. Since the pose recovery accuracy was also lower, this is in line with the observation that pose recovery accuracy affects the recognition performance.

For the HOG descriptors, the performance without CSP was lower than those obtained with the CSP classifier. However, for the HOSG descriptors, the CSP classifier performed consistently lower than the 30D pose descriptors without CSP transform. We expect that this effect was caused by the interpolation in the pose recovery step. Due to the absence of edges in the HOSG descriptor, there is probably more forward-backward ambiguity. As a result, recovered poses tend to be closer to a mean pose which has leg and arm joints close to the medial axis, thus close to the body. This mean pose is more similar to a rest pose than a common walking pose. This effect is smaller for the jog poses, as the jog action is usually performed with raised forearms. When we look at the misclassifications, we notice that most of the confusions are between the walking and rest action class. This is true both with and without using CSP transform.

6.3.2 Results under partial occlusions

In Section 3.3, we demonstrated the ability of our example-based approach to recover poses when the human figure was partly occluded. To this end, we adapted the whole descriptor normalization to normalization of each cell. Pose recovery was affected by occlusion, but only slightly. Moreover, this approach does not require retraining to cope with different occlusion settings. One of the main advantages of combining this example-based pose recovery approach with the CSP human action classifier is

that we do not have to adapt the classifier to deal with partial occlusions, as these are handled during pose recovery. In this section, we evaluate whether the pose recovery accuracy is sufficient to correctly recognize human actions under partial occlusions.

In line with Section 3.3.2, we used the *pole* sequence. This is the HumanEva-I Walking test sequence of subject 1, with a fixed part of the image occluded (see Figure 3.12). We used the approach and settings described in Section 6.2 to recover the poses inthe training set. Specifically, we normalized the whole descriptor to unit length. In informal experiments we also used normalization per cell with similar results. Here, however, we demonstrate that we can use different image representations for training and testing. The *pole* sequence contains only walking motion. Compared to the pole robustness sequence in the Weizmann human action dataset (see also Section 5.4.3), our *pole* sequence differs in that the subject walked in a circle, instead of only orthogonally to the viewing direction.

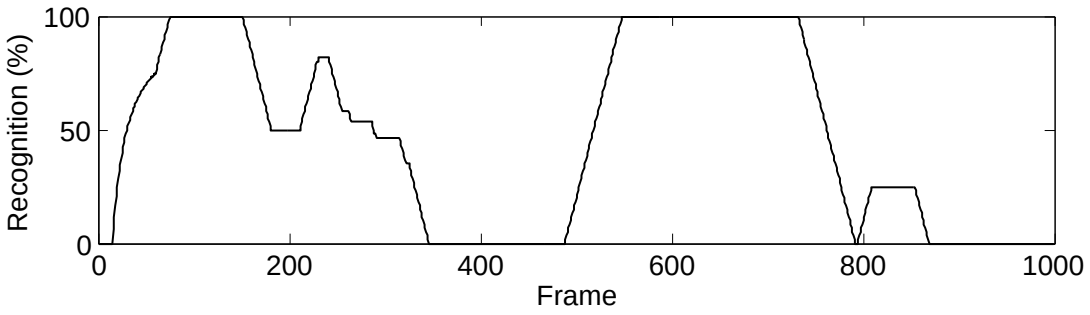
For our experiment, we again used sub-sequence lengths of 30, 60, 90 and 120 frames. Also, we used the same action class sets as in our main experiment. The walking action class is included in all of these action class sets. During testing, we used a sliding window approach with a stride length of one frame. The results are summarized in Table 6.7.

Sub-sequence length	30 frames	60 frames	90 frames	120 frames
Number of classes	15	15	7	5
CSP classifier	97.73%	100.00%	97.36%	95.68%
Without CSP (30D)	58.45%	51.38%	41.21%	66.93%
Without CSP (6D)	36.39%	0.00%	0.00%	0.00%

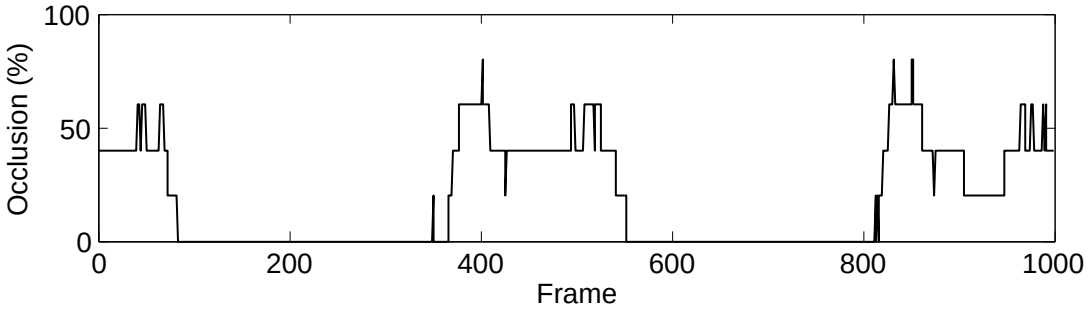
Table 6.7: Classification performance on the *pole* sequence, for different sub-sequence lengths and corresponding action subsets.

For the CSP classifier, differences between sub-sequence lengths were minimal. All confusions were between the walking and rest class. This was also true for the two settings without the CSP transform. Here, the performance was much lower, especially when only the first 6 principal components were used.

Due to the sliding window approach, each frame was included in a number of test sub-sequences. We accumulated these values for each frame. The resulting plot of these classification results without using the CSP transform, and for sub-sequence lengths of 60 frames is shown in Figure 6.3(a). The percentage of the HOG descriptor grid that is occluded is shown in Figure 6.3(b). This is the same graph as Figure 3.13(c). It is clear that there is a strong correlation between the percentage of occlusion and the action recognition performance. Specifically, when no occlusion was present, the recognition performance is generally higher than when part of the HOG descriptor is occluded. Again, we expect that the interpolation of poses results in an estimated pose that is closer to a rest pose. As a result, pose recovery accuracy will be more noisy. In contrast, the recognition performance for the CSP classifier is almost perfect. Apparently, despite this noise, the CSP classifier is able to focus on the differences between the rest and walking classes.



(a) Classification result



(b) Percentage of occlusion

Figure 6.3: Action recognition on the *pole* sequence, using a sliding window approach with sub-sequences of 60 frames, without the CSP transform. (**a**) classification results in percent per frame, averaged over all test sub-sequences which contained the frame. All confusions were with the rest action. (**b**) percentage of occlusion for the *pole* sequence.

6.3.3 Results for temporal segmentation

In the previous experiment, we used a sliding window with a fixed sub-sequence length. The sequence we used for testing contained only a single action but we can also use this approach when an image sequence contains multiple actions in succession. In this section, we evaluate the performance of this approach on several unsegmented HumanEva-I test sequences. The results that we present are qualitative, but show the potential of our approach to temporally segment sequences of human movement.

For the HumanEva-I dataset, we evaluated several test sequences that contain segments of multiple action classes. Most of these include action classes that we could not recognize because we did not have sufficient training data. This was especially true for the Throw/Catch and Box activities, and for the 'free-style' part of the Combo sequences. Even though we could not obtain a correct classification for these segments, we included them in our evaluation as it gives a better idea of how well our approach performs on the segmentation of unseen actions. Specifically, we evaluated the Gesture, Throw/Catch and Combo sequences of subject 1. Also, some of the segments that we evaluated were really short. In our previous experiments, these were left out but we included them in our segmentation experiment.

We processed the sequences with the sliding window approach with a stride length

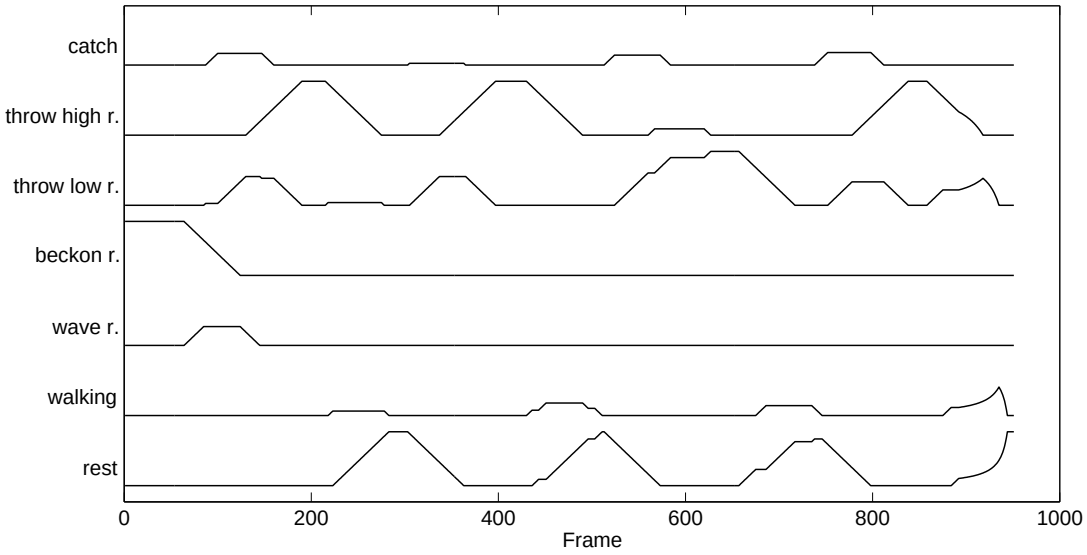


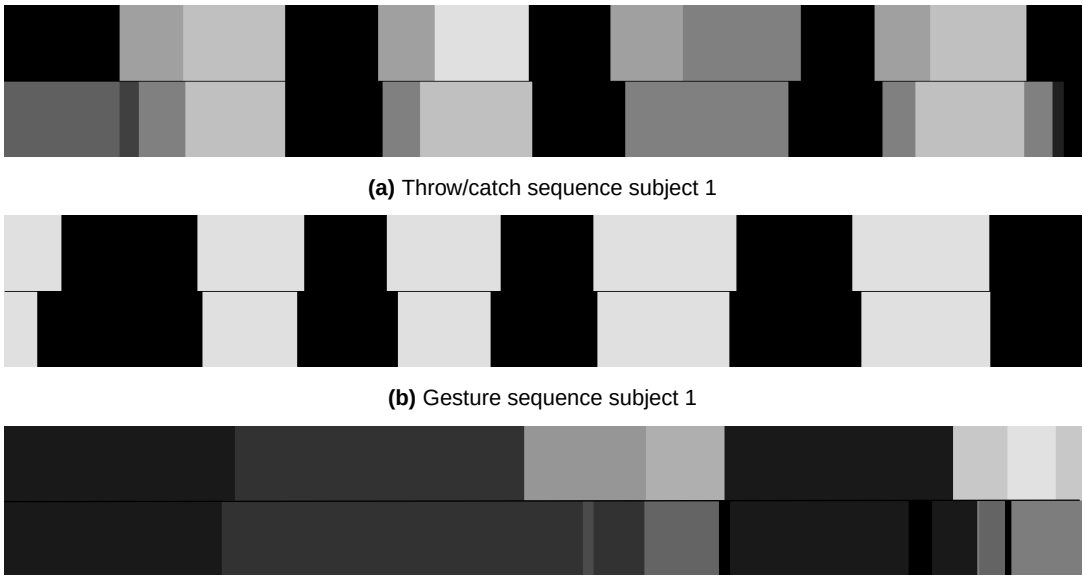
Figure 6.4: Estimated action recognition probabilities for the Throw/catch sequence of subject 1.

of one frame. Sub-sequences of 60 frames were evaluated, using the CSP classifier with the settings of the experiment in Section 6.2. Specifically, we used the HOG-F-5×6 descriptors and 15 action classes. For each frame, the estimated class label was averaged over all sub-sequences the frame was in. Figure 6.4 shows an example where all action classes that were not guessed are omitted. These values could be interpreted as the probabilities that a frame is part of a certain action segment, given all the sub-sequence classifications. To determine the segmentation we calculated the most common classification label for each frame.

Figure 6.4 shows the recognition probabilities for the Throw/Catch sequence of subject 1. It is clear that the rest and the low and high throw action classes were guessed several times. There were also some incidental estimates for gesture actions and walking. Next, we looked at the action class that had the highest probability. Figure 6.5(a) shows this segmentation, with the ground truth at the top and the obtained segmentation at the bottom. For the Gesture and Combo sequences, these segmentations appear in Figures 6.5(b) and (c), respectively. In each plot, similar colors correspond to similar actions. Colors between plots may correspond to different action classes. There are a considerable number of misclassifications in the Throw/Catch and Combo sequences. Most of these were due to the fact that the correct action labels were not part of the training set. The Gesture sequence contains only wave and beckon segments, performed with the right hand.

Overall, there is a clear correlation between the boundaries between segments in the ground truth and those obtained from our action recognition approach. However, there are differences. For example, in Figure 6.5(a), the black segments correspond to the rest action. It can be seen that our approach classified more frames as rest. A similar observation can be made for the black segments in Figure 6.5(b), which correspond to the wave action. Such differences are common, as is is not clear where one action ends and the next one starts. For the second half of the Combo sequence,

the segmentation of our approach differs somewhat from the ground truth. This is due to the fact that none of the balancing and jumping actions were in the training set.



(a) Throw/catch sequence subject 1

(b) Gesture sequence subject 1

(c) Combo sequence subject 1

Figure 6.5: Action segmentation for three HumanEva-I test sequences of subject 1. The upper row in each plot corresponds to the ground truth, the bottom row is the segmentation that is obtained from our action recognition approach using a sliding window.

6.4 Discussion

In this section, we have combined the example-based pose recovery approach with the CSP classifier to recognize human actions using recovered poses. Combining these two algorithms solves several challenges that are difficult to address in a single step, such as dealing with partial occlusions and variations in viewpoint. Moreover, our combined approach has the advantage that action models can be trained using motion capture data, if the motion capture data and pose recovery adhere to a similar pose representation. We have evaluated our approach on the HumanEva-I dataset, where segments of movement have been labeled with action class labels. We have performed experiments with different sub-sequence lengths. Also, we have compared the performance of the CSP classifier with the performance without using a CSP transform. Using CSP proved to be advantageous, both in recognition performance and in the number of dimensions that was used to describe the action prototype vectors. For sub-sequences of 60 frames (1 second), performance for the CSP classifier was 87.20%, whereas classification without CSP on PCA-reduced pose vectors of the same dimensionality was only 65.46%. These performances are largely due to a majority of samples of the jog and walking classes, which have often been classified correctly. This shows that our approach can distinguish between related classes. It is to be

expected that performance is to increase for the other classes when more training examples are available.

By normalizing the poses for viewpoint and body dimensions, we could train and test our classifier with data obtained from different viewpoints and subjects. Our experiments showed that we could recognize walking and jog motions reliably even when they were performed in a circle.

Previously, we demonstrated that we could recover poses under partial occlusion if the occlusion regions were labeled. In this section, we have shown that we could recognize walking movements without adapting the training of the CSP classifier. Recognition performance was unaffected when using the CSP classifier but decreased significantly when no CSP transform was used.

By using a sliding window approach, we demonstrated that the classifier could also be used to temporally segment a sequence into action instances.

Despite the advantages of combining the example-based pose recovery approach with the CSP classifier, several issues remain. First of all, to be able to recognize human action from arbitrary viewpoints, these poses must be recovered first. This requires sufficient close examples of the pose from a specific viewpoint. One potential solution is to learn the mapping using synthetically generated image representations from recorded motion capture data. As such, our approach is more versatile than most human action recognition approaches that directly rely on image data.

Also, there are still several issues in the classification stage that could be addressed. First, by representing an entire class by a single low-dimensional prototype vector, we ignored the intra-class variance. Also, these action prototypes do not take into account temporal characteristics. Second, our approach requires that it is trained on all types of action that it is to observe. This means that unseen action types will be classified as one of the known action classes. We could adapt our approach to include an 'other' class but it is generally difficult to describe such a class due to the large variation of movements that could be part of it.

Part III

Conclusion

7

Discussion and future research

This chapter concludes this thesis. In Section 7.1, we summarize our contribution to the field of human pose recovery and action recognition. The strengths and limitations of our contributions are discussed in Section 7.2. Finally, we present avenues for future work in the domains of vision-based recovery and recognition of human motion in Section 7.1.

7.1 Summary of our contribution

We briefly summarize the contributions that we made in this thesis. A detailed discussion of the strengths and limitations of our contributions appears in Section 7.2.

- In Chapters 2 and 4, we presented an extensive overview of literature in human pose recovery and human action recognition. We discussed the relative (dis)advantages of different approaches, and pointed out limitations in the current state of the art.

- In Chapter 3, we introduced our example-based human pose recovery approach where a variant of histograms of oriented gradients (HOG) was used as the image descriptor. Given an unseen image of a subject, represented as a HOG descriptor, we used weighted interpolation between the visually closest examples of an example dataset to obtain the pose estimate. This approach is fast due to the use of a low-cost distance function (Manhattan distance). We obtained reasonably accurate results on the publicly available HumanEva dataset. In our experiments, we examined the effect on the accuracy when varying the HOG grid size and the number of examples in the example set.

- In Section 3.3, we adapted our example-based human pose recovery approach to cope with partial occlusion of the human subject in the image. The normalization and matching of the HOG descriptors was changed from global to the cell level. The approach was used for online recovery of human poses without adjusting the example set and matching process to the occlusion condition.

Experiments on the HumanEva dataset with simulated occlusion showed that occlusions affected the recovery accuracy but only moderately.

- We presented a human action recognition approach using common spatial patterns (CSP) in Chapter 5. Simple functions were used to discriminate between two classes by transforming sequences of HOG descriptors. In the transformation the difference in variance between the two classes was maximized. Each of the discriminative functions softly voted into two classes. After evaluation of all pairwise functions, the action class that received most of the voting mass was the estimated class. Experiments on the publicly available Weizmann human action dataset showed that we could obtain state of the art performance with low training requirements. We showed that walking motion could be recognized even for moderate viewpoint changes and with several image deformations. Also, we obtained good results when shorter sequences were used.

- In Chapter 6, we combined the example-based pose recovery approach with the CSP-based human action recognition approach. Specifically, we used sequences of recovered poses as input for the CSP classifier. Again, we used the publicly available HumanEva dataset for our experiments. Thanks to the rotation normalization, actions could be recognized from arbitrary viewpoints. Moreover, by handling occlusions in the pose recovery step, we could recognize actions from partially occluded image observations. We further demonstrated the potential of our approach for temporal action segmentation.

7.2 Discussion of our approach

Our contributions have several advantages over existing work as we will discuss in this section. Also, there are some limitations to our approach which we will explain as well. We discuss human pose recovery and human action recognition separately in Sections 7.2.2 and 7.2.3, respectively. Since the use of the HOG and HOSG descriptors is common in both of these topics, we discuss their use separately in Section 7.2.1. Finally, in Section 7.3, we present some directions of future research to address some of the current limitations.

7.2.1 Image descriptors

Both for the recovery of human poses and the recognition of human actions, we encoded the image observations using a variant of histograms of oriented gradients (HOG). This is a grid-based descriptor that uses edge orientation histograms. We adapted the originally proposed HOG by removing a Gaussian edge smoothing step, and by discarding the notion of overlapping blocks of cells. Most importantly, we only took into account those edges that were part of the extracted foreground region. As such, the descriptor only focusses on the subject, and ignores edges in the background. This makes the descriptor much more informative. In addition, we introduced histograms of oriented silhouette gradients (HOSG), which did not use the edge information but only took into account the foreground silhouette boundary points. Calculation of the HOG/HOSG descriptor could be performed efficiently.

In general, differences in performance between HOG and HOSG descriptors were small. Pose recovery accuracy was slightly higher when edges where used. This probably reduced the number of forward-backward ambiguities. For action recognition, we obtained slightly higher recognition rates when using the HOSG descriptors. As we only considered fronto-parallel movement, forward-backward ambiguities were rare. The additional edge information of the HOG descriptors also encoded the appearance and clothing of the subject. This information is not important for subject-invariant action recognition and is a limitation of the image descriptor.

For both the HOG descriptor and the HOSG descriptor, we relied on background subtraction. Also, we required that the region of interest (ROI) was known. We determined the ROI from the extracted foreground region. Our approach is therefore dependant on relatively accurate background segmentation. In our experiments, this segmentation was reasonable, but we only used video sequences with relatively static backgrounds. For more complex environments, it might prove more difficult to obtain accurate foreground segmentation. Also, this might require more complex background segmentation algorithms, which are computationally more demanding.

If the global descriptor normalization is changed to a cell-based normalization, we can cope with partial occlusions of the human figure in the image. The only requirement is that the occlusion area must be known. Due to the grid-based nature of the HOG and HOSG descriptors, cells where occlusion occurs can be ignored in the matching. Consequently, there is no need to change the example set to cope with different occlusion settings. Experiments on the HumanEva dataset with simulated occlusion demonstrated that pose recovery accuracy decreased with an increasing percentage of occlusion. However, this decrease was relatively small, compared to the percentage of occlusion.

7.2.2 Human pose recovery

We took a discriminative approach to human pose recovery. This implies that no iterative model matching stage is needed to recover the pose from an image observation. Consequently, our work is fast and can be made to run in real time is human detection and segmentation can be performed efficiently. The main limitation of a discriminative approach is that only poses can be recovered that are within the convex hull of all training examples. Moreover, the training examples have to cover this pose space relatively densely, as our experiment with different numbers of examples indicates. Our approach is therefore suitable for situations where speed of processing is important, but the pose space is limited. Human-computer interaction and surveillance and security are typical application domains where our contributions are valuable. In general, discriminative approaches are less accurate as they are not suitable to explicitly recover body dimension parameters. A generative approach is therefore more suitable for applications that require precise recovery of body pose. However, a combination in which a discriminative approach is used to find initial pose estimates can significantly speed up the processing.

We used an example-based approach where we retained all training examples. Since we need to find the visually most similar examples, the computational complexity of recovering a pose is linear in the number of training examples. More examples

are required for broader pose domains. This will decrease the computational speed and increase memory requirements to store all examples. Instead, a regression-based approach may be used. We will discuss this in the next section.

Generative approaches usually require an approximate initial estimation of the pose. In general, such an estimate is obtained through tracking of the pose over time. Still, this requires pose initialization of the first frame. In contrast, our approach does not require initialization and can be used to recover the pose from a single frame. We did not model the relation between subsequent frames or subsequent poses. In theory, the use of such a dynamical model could reduce the number of pose ambiguities but presents the risk of being too restrictive. Moreover, application of a dynamical model again requires good initialization.

In our approach, the examples in the database should also cover the range of viewpoints that we expect to observe in testing conditions. This is the result of the use of HOG/HOSG descriptors, which are not viewpoint-invariant. Even though we performed experiments with observations from multiple views, we did not combine these into a single 3D observation. In our experiments, the multi-view descriptors proved to be slightly more accurate, but at the cost of a tripled descriptor length and the limitation that the camera setup of training and test data had to be identical.

7.2.3 Human action recognition

Our common spatial patterns (CSP) classifier for recognition of human actions makes use of simple functions that discriminate between pairs of classes. Evaluation of these functions can be performed real-time. Moreover, training requirements are low. Construction of the discriminative functions has a low computational complexity. Moreover, we have shown that our approach obtained good results when training on data of only a few of the available training subjects. This indicates that generalization between subjects was high. Also, when adding new actions, there is no need to retrain all discriminative functions.

In our CSP classifier, each action class was represented by a single low-dimensional vector. Such a compact representation allows for fast evaluation of the discriminative functions. However, the representation is naive as no intra-class variance is modeled. Therefore, such a representation is likely to be too simple, especially for classes that can vary significantly in performance, for example when there are multiple ‘modes’ of movement. Moreover, we did not explicitly model the temporal aspect within the performance of the movement. Consequently, our approach is unable to distinguish between movements that have a different order of performance. Despite these limitations, our experiments showed that we could achieve state of the art performance on the publicly available Weizmann human action dataset.

Our approach has the limitation that, given a sequence of movement, one of the trained classes is always guessed. There is no mechanism to determine whether the observed movement belongs to neither of the classes. This forced-choice approach makes the application less suitable for more realistic domains. To overcome this limitation, an ‘other’ class could be added. We discuss this in the next section.

We also investigated how well we could recognize actions from recovered poses. To this end, we combined our example-based pose recovery approach with the CSP

classifier. The recovered poses were normalized for rotations around a vertical axis and normalized for body height. Pose descriptors proved to be compact, yet sufficiently rich to distinguish between several human action classes. One of the advantages of our combined approach is that the classifier can be trained using motion capture data. Moreover, the trained models are viewpoint invariant. We used the HumanEva dataset for our experiments. In this set, there is considerable variation in the performance of actions between subjects. Also, the direction of movement varied, especially in the walking and jog sequences, where the movement was performed in a circle. The almost perfect recognition of these motions in our experiments demonstrates the ability of our approach to cope with viewpoint changes.

Another advantage is that we were able to recognize walking motion even though part of the observation was occluded. These occlusions were handled in the pose recovery approach, and did not affect recognition performance. No change in the training or evaluation steps was required to cope with these occlusions.

A drawback of our combined human action recognition approach is that we depend on pose recovery. This implies that we are only able to recognize actions from viewpoints that are included in the example set. Also, several of the drawbacks of the CSP classifier remain. Specifically, we did not model intra-class variance. Also, temporal information which a sequence over time was not taken into account.

Additional experiments showed the potential of our approach to perform simultaneously human action recognition and temporal segmentation of sequences into action segments. In these experiments, not all action classes were known and it is to be expected that both segmentation and recognition performance increase with a more complete repertoire of action classes.

7.3 Future research

There are a number of avenues for future research which we present in this section. We discuss potential directions in human pose recovery and human action recognition in Sections 7.3.1 and 7.3.2. We have considered the evaluation of our work as a vital part of our contribution. Therefore, in Section 7.3.3, we discuss several ideas that could leverage the quality of current evaluation practice.

7.3.1 Human pose recovery

Our pose recovery approach is focussed on real-time applications. We used an example-based approach, mainly because there are no parameters that required tuning. However, the matching cost of a HOG descriptor of an unseen frame with the example set is linear in the number of examples. Larger example sets, which are required when more activities and viewpoints are considered, will severely decrease recovery speed. Hashing (such as in [310]) or hierarchical matching (for example [109]) will significantly speed up the matching process.

Alternatively, a regression-based approach could be adopted (e.g. [80; 329]). This significantly reduces matching time and scales favorably to more unconstrained pose spaces. However, as gating and regression functions are usually trained simultaneously, training can become prohibitively time-consuming when more examples are

available. Recent work by Bo *et al*. [33] addresses this problem by training the gating functions and regression functions separately.

Another important issue is the choice of image descriptor. In our experiments, we have used HOG and HOSG descriptors. For pose recovery, the edges in the HOG descriptor appeared to be advantageous to avoid forward-backward ambiguities. However, it appeared that the use of both HOG and HOSG descriptors was somewhat person-specific. This severely degrades generalization and makes training of the regressors difficult. Okada and Soatto [253] discriminatively select those image features that are informative of the pose. In related work by Kanaujia *et al*. [164], image observations are encoded hierarchically at multiple levels. Both approaches reduce the need for accurate detection and background subtraction. Still, accurate localization within the image will drastically reduce computation time. It is to be expected that recent work in the area of 3D object recognition will continue to be a good starting point for adaptation to the human detection task.

Recovering poses from partially occluded image observations has been an ignored topic within the domain. Our approach is one of the first to address the issue while focussing on real-time applications. To cope with partial occlusions, our approach required predicted occlusion. While it is generally difficult to obtain such a labelling, the field of human detection is increasingly focussing on dealing with partial occlusions. Recent work by Lin and Davis [198] and Wu and Nevatia [394] addresses this issue, but it remains challenging to reliably detect humans and identify occlusion regions in unconstrained domains. More research is required to be able to perform reliable detection and occlusion prediction in more complex situations.

For our example set, we used poses with their corresponding image observations, obtained from synchronized video sequences. In general, it is costly to record this type of data, especially considering the fact that only a limited number of viewpoints can be used. To overcome this problem, synthetic image observations can be generated using 3D animation software. This approach has been successfully used in [4; 81; 327].

In our approach, we focussed on recovery from a single frame. However, when a sequence of frames is available, improved recovery accuracy can be obtained by considering temporal consistency in the poses. This could be enforced by employing a dynamic model. Not only would such an approach ensure that movement over time is realistic, the search for close matches can also be more focussed on those examples that have poses close to the expected pose. However, applying such a model has the drawback that it should be properly initialized which is generally a difficult task.

Finally, a combination of generative and discriminative approaches might benefit from the advantages of either [98; 283; 320]. Fast initial estimates can be obtained from the discriminative part. This also solves the initialization issue in generative approaches. The generative part allows for further refinement of this pose.

7.3.2 Human action recognition

In our human action recognition approach, both when using image descriptors and when using recovered poses, an action class was represented by a low-dimensional vector. Such a representation is compact, but does not allow for modeling of the temporal structure of the action. Nor does it model intra-class variance. A more

complex representation of each action would allow for better recognition, especially for actions that have more variety in performance.

Our approach used a forced-choice classification scheme. When actions can be observed that are not part of the trained set of actions, such a scheme will lead to misclassifications. Therefore, a mechanism is required to detect unknown classes. One approach is to include an 'other' class which competes against all other classes. However, such a class will contain much variation and should be trained on examples that are not in the training set. This makes it cumbersome to use such a model. Alternatively, a threshold on the voting mass that is attributed to a certain class can be used. In our approach, we selected the action class that received most of the voting mass, regardless of the value. By setting a threshold, or by looking at differences between the highest scoring classes, a more informed decision could be made.

When recovering poses, knowledge about the action that is being is performed could result in increased recovery accuracy. In addition, when recognizing human action, we have shown that poses are suitable descriptors. In our combined approach, we used recovered poses for classification. Another approach would be to link the two tasks, for example using a graphical model [44; 119; 268]. However, it is an open challenge for these approaches to deal with unseen action classes.

7.3.3 Evaluation practice

For the evaluation of our approaches, we have used publicly available databases. These allow for objective comparison between different algorithms. Moreover, since the challenges of each dataset are known, the strengths and limitations of an approach can better be understood. However, the use of these datasets presents the risk of tailoring the approach to the dataset. This might lead to improved results, but is likely to limit generalization ability. Therefore, there is the ongoing need for more challenging datasets.

With the progressing sophistication of human pose recovery and action recognition approaches, the application of these algorithms in realistic scenarios comes within reach. These scenarios should address occlusions, more variation in activities and a broader variety of viewpoints. The recovery and recognition of human movement has applications in several domains such as surveillance, human-computer interaction and video retrieval. The requirements for these domains are different. Human-computer interaction applications require real-time processing, missed detections in surveillance are unacceptable and video retrieval applications often cannot benefit from a controlled setting. Given these differences, it seems reasonable to record different datasets for the various domains. This would allow for the use of evaluation metrics that go beyond precision and recall measures, such as speed of processing or detection accuracy.

Given the current state of the art, and motivated by the broad range of applications that can benefit from robust human pose recovery and action recognition, it is to be expected that many of the challenges will be addressed in the near future. This would open the door to fulfill the longstanding goal of robust automatic interpretation of human motion.

Bibliography

[1] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram. A novel approach for recognition of human actions with semi-global features. *Machine Vision and Applications*, 19(1):27–34, January 2008.

[2] Ankur Agarwal and Bill Triggs. Tracking articulated motion using a mixture of autoregressive models. In *Proceedings of the European Conference on Computer Vision (ECCV'04) - volume 3*, number 3024 in Lecture Notes in Computer Science, pages 54–65, Prague, Czech Republic, May 2004.

[3] Ankur Agarwal and Bill Triggs. A local basis representation for estimating human pose from cluttered images. In *Proceedings of the Asian Conference on Computer Vision (ACCV'06) - part 1*, number 3851 in Lecture Notes in Computer Science, pages 50–59, Hyderabad, India, January 2006.

[4] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(1):44–58, January 2006.

[5] Jake K. Aggarwal and Qin Cai. Human motion analysis: a review. *Computer Vision and Image Understanding (CVIU)*, 73(3):428–440, March 1999.

[6] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3):1–10, August 2008.

[7] Edilson de Aguiar, Christian Theobalt, Carsten Stoll, and Hans-Peter Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[8] Md. Atiqur Rahman Ahad, Takehito Ogata, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Motion recognition approach to solve overwriting in complex actions. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08)*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[9] Mohiuddin Ahmad and Seong-Whan Lee. Human action recognition using shape and CLG-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, July 2008.

[10] Saad Ali, Arslan Basharat, and Mubarak Shah. Chaotic invariants for human action recognition. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[11] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear, 2009.

[12] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *The Journal of Machine Learning Research*, 1:113–141, December 2000.

[13] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, July 2005.

[14] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1–5):11–73, February 1997.

[15] Alexandru O. Bălan and Michael J. Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 2*, number 5303 in Lecture Notes in Computer Science, pages 15–29, Marseille, France, October 2008.

[16] Alexandru O. Bălan, Michael J. Black, Horst Haussecker, and Leonid Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[17] Alexandru O. Bălan, Leonid Sigal, and Michael J. Black. A quantitative evaluation of video-based 3D person tracking. In *Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05)*, pages 349–356, Beijing, China, October 2005.

[18] Alexandru O. Bălan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[19] Jan Bandouch, Florian Engstler, and Michael Beetz. Evaluation of hierarchical sampling strategies in 3D human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 925–934, Leeds, United Kingdom, September 2008.

[20] Carlos Barrón and Ioannis A. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding (CVIU)*, 81(3):269–284, March 2001.

[21] Carlos Barrón and Ioannis A. Kakadiaris. Monocular human motion tracking. *Multimedia Systems*, 10:118–130, October 2004.

[22] Dhruv Batra, Tsuhan Chen, and Rahul Sukthankar. Space-time shapelets for action recognition. In *Proceedings of the Workshop on Motion and Video Computing (WMVC'08)*, pages 1–6, Copper Mountain, CO, January 2008.

[23] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(4):509–522, January 2002.

[24] Chiraz BenAbdelkader and Larry S. Davis. Estimation of anthropomeasures from a single calibrated camera. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'06)*, pages 499–504, Southampton, United Kingdom, April 2006.

[25] Michael van den Bergh, Esther Koller-Meier, and Luc J. van Gool. Real-time body pose recognition using 2D or 3D haarlets. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[26] Christopher Bishop and Markus Svensén. Bayesian hierarchical mixtures of experts.

In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'03)*, pages 57–64, Acapulco, Mexico, April 2003.

[27] Alessandro Bissacco, Ming-Hsuan Yang, and Stefano Soatto. Detecting humans via their pose. In *Advances in Neural Information Processing Systems (NIPS) 19*, pages 169–176, Vancouver, Canada, December 2006.

[28] Alessandro Bissacco, Ming-Hsuan Yang, and Stefano Soatto. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[29] Jaron Blackburn and Eraldo Ribeiro. Human motion recognition using Isomap and dynamic time warping. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 285–298, Rio de Janeiro, Brazil, October 2007.

[30] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 2*, pages 1395–1402, Beijing, China, October 2005.

[31] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, December 1997.

[32] Liefeng Bo and Cristian Sminchisescu. Twin Gaussian processes for structured prediction. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[33] Liefeng Bo, Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Fast algorithms for large scale conditional 3D prediction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[34] Aaron F. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1257–1265, August 1997.

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

[36] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision (IJCV)*, 74(1):17–31, August 2007.

[37] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR'07)*, pages 401–408, Amsterdam, The Netherlands, July 2007.

[38] Andrea Bottino and Aldo Laurentini. A silhouette-based technique for the reconstruction of human movement. *Computer Vision and Image Understanding (CVIU)*, 83(1):79–95, July 2001.

[39] Richard Bowden, Tom A. Mitchell, and Mansoor Sarhadi. Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences. *Image and Vision Computing*, 18(9):729–737, June 2000.

[40] Matthew Brand. Shadow puppetry. In *Proceedings of the International Conference on Computer Vision (ICCV'99) - volume 2*, pages 1237–1244, Kerkyra, Greece, September 1999.

[41] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and

tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3):179–194, February 2004.

[42] Marcus A. Brubaker and David J. Fleet. The kneed walker for human pose tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[43] Marcus A. Brubaker, David J. Fleet, and Aaron Hertzmann. Physics-based person tracking using simplified lower-body dynamics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[44] Fabrice Caillette, Aphrodite Galata, and Toby Howard. Real-time 3-D human body tracking using learnt models of behaviour. *Computer Vision and Image Understanding (CVIU)*, 109(2):112–125, February 2008.

[45] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Transactions on Computer Graphics*, 22(3):569–577, July 2003.

[46] Bhaskar Chakraborty, Ognjen Rudovic, and Jordi Gonzàlez. View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08)*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[47] Tat-Jen Cham and James M. Rehg. A multiple hypothesis approach to figure tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'99) - volume 2*, pages 239–245, Ft. Collins, CO, June 1999.

[48] Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen, and Suh-Yin Lee. Human action recognition using star skeleton. In *Proceedings of the International Workshop on Video Surveillance and Sensor Networks (VSSN'06)*, pages 171–178, Santa Barbara, CA, October 2006.

[49] Shinko Y. Cheng and Mohan M. Trivedi. Articulated human body pose inference from voxel data using a kinematically constrained Gaussian mixture model. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (CVPR-EHuM)*, Minneapolis, MN, June 2007.

[50] Srikanth Cherla, Kaustubh Kulkarni, Amit Kale, and Viswanathan Ramasubramanian. Towards fast, view-invariant human action recognition. In *Proceedings of the Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB'08)*, pages 1–8, Anchorage, AK, June 2008.

[51] German K. Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03) - volume 1*, pages 77–84, Madison, WI, June 2003.

[52] Tat-Jun Chin, Liang Wang, Konrad Schindler, and David Suter. Extrapolating learned manifolds for human activity recognition. In *Proceedings of the International Conference on Image Processing (ICIP'07) - volume 1*, pages 381–384, San Antonio, TX, September 2007.

[53] Olivier Chomat, Jérôme Martin, and James L. Crowley. A probabilistic sensor for the perception and recognition of activities. In *Proceedings of the European Conference on Computer Vision (ECCV'00) - volume 1*, number 1842 in Lecture Notes in Computer Science, pages 487–503, Dublin, Ireland, June 2000.

[54] Chi-Wei Chu, Odest C. Jenkins, and Maja J. Mataric. Markerless kinematic model and motion capture from volume sequences. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03) - volume 2*, pages 475–483, Madison, WI, June 2003.

[55] Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4*, number 5305 in Lecture Notes in Computer Science, pages 158–171, Marseille, France, October 2008.

[56] Ross Cutler and Larry S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):781–796, August 2000.

[57] Fabio Cuzzolin. Using bilinear models for view-invariant action and identity recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1701–1708, New York, NY, June 2006.

[58] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1*, pages 886–893, San Diego, CA, June 2005.

[59] Somayeh Danafar and Niloofar Gheissari. Action recognition for surveillance applications using optic flow and SVM. In *Proceedings of the Asian Conference on Computer Vision (ACCV'07) - part 2*, number 4844 in Lecture Notes in Computer Science, pages 457–466, Tokyo, Japan, November 2007.

[60] John Darby, Baihua Li, and Nicholas Costen. Behaviour based particle filtering for human articulated motion tracking. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[61] Ben Daubney, David Gibson, and Neill Campbell. Real-time pose estimation of articulated objects using low-level motion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[62] Yiğithan Dedeoğlu, B. Uğur Töreyin, Uğur Güdükbay, and A. Enis Çetin. Silhouette-based method for object classification and human action recognition in video. In *Proceedings of the Workshop on Computer Vision in Human-Computer Interaction (ECCV-HCI'06)*, number 3979 in Lecture Notes in Computer Science, pages 64–77, Graz, Austria, May 2007.

[63] Quentin Delamarre and Olivier Faugeras. 3D articulated models and multiview tracking with physical forces. *Computer Vision and Image Understanding (CVIU)*, 81(3):328–357, March 2001.

[64] David Demirdjian and Raquel Urtasun. Patch-based pose inference with a mixture of density estimators. In *Proceedings of the International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'07)*, number 4778 in Lecture Notes in Computer Science, pages 96–108, Rio de Janeiro, Brazil, October 2007.

[65] Arthur P. Dempster, Nam M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, January 1977.

[66] Jonathan Deutscher and Ian Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2):185–205, February 2005.

[67] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286,

January 1995.

[68] David E. DiFranco, Tat-Jen Cham, and James M. Rehg. Reconstruction of 3-D figure motion from 2-D correspondences. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01) - volume 1*, pages 307–314, Kauai, HI, December 2001.

[69] Miodrag Dimitrijevic, Vincent Lepetit, and Pascal Fua. Human body pose detection using Bayesian spatio-temporal templates. *Computer Vision and Image Understanding (CVIU)*, 104(2–3):127–139, November 2006.

[70] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05)*, pages 65–72, Beijing, China, October 2005.

[71] Lan Dong, Vasu Parameswaran, Visvanathan Ramesh, and Imad Zoghlami. Fast crowd segmentation using shape indexing. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[72] Tom Drummond and Roberto Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *Proceedings of the International Conference On Computer Vision (ICCV'01) - volume 2*, pages 315–320, Vancouver, Canada, July 2001.

[73] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 2*, pages 726–733, Nice, France, October 2003.

[74] Carl Henrik Ek, Jonathan Rihan, Philip H.S. Torr, Grégory Rogez, and Neil D. Lawrence. Ambiguity modeling in latent spaces. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (MLMI'08)*, number 5237 in Lecture Notes in Computer Science, pages 62–73, Utrecht, The Netherlands, September 2008. Springer-Verlag.

[75] Carl Henrik Ek, Philip H.S. Torr, and Neil D. Lawrence. Gaussian process latent variable models for human pose estimation. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (MLMI'07)*, number 4892 in Lecture Notes in Computer Science, pages 132–143, Brno, Czech Republic, June 2007. Springer-Verlag.

[76] Ahmed M. Elgammal and Larry S. Davis. Probabilistic framework for segmenting people under occlusion. In *Proceedings of the International Conference On Computer Vision (ICCV'01) - volume 2*, pages 145–152, Vancouver, Canada, July 2001.

[77] Ahmed M. Elgammal and Chan-Su Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 681–688, Washington, DC, June 2004.

[78] Ahmed M. Elgammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 1*, pages 478–485, Washington, DC, June 2004.

[79] Ahmed M. Elgammal and Chan-Su Lee. Nonlinear manifold learning for dynamic shape and dynamic appearance. *Computer Vision and Image Understanding (CVIU)*, 106(1):31–46, April 2007.

[80] Ahmed M. Elgammal and Chan-Su Lee. Tracking people on a torus. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(3):520–538, March 2009.

[81] Markus Enzweiler and Dariu M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[82] Markus Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear, 2009.

[83] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):52–73, October 2007.

[84] María-José Escobar, Guillaume S. Masson, Thierry Vieville, and Pierre Kornprobst. Action recognition using a bio-inspired feedforward spiking network. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[85] Claudio Fanti, Lihi Zelnik-Manor, and Pietro Perona. Hybrid models for human motion recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1*, pages 1166–1173, San Diego, CA, June 2005.

[86] Ali Farhadi and Mostafa Kamali Tabriz. Learning to recognize activities from the wrong view point. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 1*, number 5302 in Lecture Notes in Computer Science, pages 154–166, Marseille, France, October 2008.

[87] Alireza Fathi and Greg Mori. Human pose estimation using motion exemplars. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, Ocotber 2007.

[88] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[89] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

[90] Pedro F. Felzenszwalb, David McAllester, and Deva Ramanan. Discriminatively trained multiscale deformable part models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[91] Tien-Chieng Feng, Prabath Gunawardane, and James Davis Bolan Jiang. Motion capture data retrieval using an artists doll. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[92] Xiaolin Feng and Pietro Perona. Human action recognition by sequence of movelet codewords. In *Proceedings of the International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'02)*, pages 717–721, Padova, Italy, June 2002.

[93] Vittorio Ferrari, Manuel Marín-Jiménez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[94] Preben Fihl, Michael B. Holte, Thomas B. Moeslund, and Lars Reng. Action recognition using motion primitives and probabilistic edit distance. In *International Workshop on Articulated Motion and Deformable Objects (AMDO'06)*, number 4069 in Lecture Notes in Computer Science, pages 375–384, Port d'Andratx, Spain, July 2006.

[95] Roman Filipovych and Eraldo Ribeiro. Learning human motion models from unsegmented videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–7, Anchorage, AK, June 2008.

[96] Mathias Fontmarty, Frédéric Lerasle, and Patrick Danès. Towards real-time markerless human motion capture from ambiance cameras using an hybrid particle filter. In *Proceedings of the International Conference on Image Processing (ICIP'08)*, pages 709–712, San Diego, CA, October 2008.

[97] 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):77–254, July 2006.

[98] Andrea Fossati, Elise Arnaud, Radu Horaud, and Pascal Fua. Tracking articulated bodies using generalized expectation maximization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pages 1–6, Washington, DC, June 2008.

[99] Andrea Fossati, Miodrag Dimitrijevic, Vincent Lepetit, and Pascal Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[100] Andrea Fossati and Pascal Fua. Linking pose and motion. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4*, number 5305 in Lecture Notes in Computer Science, pages 200–213, Marseille, France, October 2008.

[101] William T. Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *Proceedings of the Workshop on Automatic Face and Gesture Recognition*, pages 296–301, Zurich, Switzerland, June 1995.

[102] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[103] Jerry H. Friedman. Another approach to polychotomous classification. Statistics department, Stanford University, Stanford, CA, October 1996.

[104] Jürgen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[105] Jürgen Gall, Bodo Rosenhahn, and Hans-Peter Seidel. Clustered stochastic optimization for object recognition and pose estimation. In *Proceedings of the Symposium on Pattern Recognition (DAGM'06)*, number 4713 in Lecture Notes in Computer Science, pages 32–41, Heidelberg, Germany, September 2007.

[106] Tarak Gandhi and Mohan M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions On Intelligent Transportation Systems*, 8(3):413–430, September 2007.

[107] Tarak Gandhi and Mohan M. Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *Proceedings of the Intelligent Vehicles Symposium (IV'08)*, pages 506–511, Eindhoven, The Netherlands, June 2008.

[108] Dariu M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–92, January 1999.

[109] Dariu M. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(8):1408–1421, August 2007.

[110] Dariu M. Gavrila and Larry S. Davis. Tracking of humans in action: A 3D model-based approach. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*

(CVPR'96), pages 73–80, San Francisco, CA, June 1996.

[111] Andrew Gilbert, John Illingworth, and Richard Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 1*, number 5302 in Lecture Notes in Computer Science, pages 222–233, Marseille, France, October 2008.

[112] Laetitia Gond, Patrick Sayd, Thierry Chateau, and Michel Dhome. A 3D shape descriptor for human pose recovery. In *International Workshop on Articulated Motion and Deformable Objects (AMDO'08)*, number 5098 in Lecture Notes in Computer Science, pages 370–379, Port d'Andratx, Spain, July 2008.

[113] Neil J. Gordon, David J. Salmond, and Adrian F.M. Smith. Novel approach to nonlinear/nongaussian Bayesian state estimation. In *IEE Proceedings-F (Radar and Signal Processing)*, volume 140, pages 107–113, April 1993.

[114] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(12):2247–2253, January 2007.

[115] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 1*, pages 641–647, Nice, France, October 2003.

[116] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popovic. Style-based inverse kinematics. *ACM Transactions on Graphics*, 23(3):522–531, August 2004.

[117] Matthias Grundmann, Franziska Meier, and Irfan Essa. 3D shape context and distance transform for action recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[118] Feng Guo and Gang Qian. Human pose inference from stereo cameras. In *Proceedings of the Workshop on Applications of Computer Vision (WACV'07)*, page 37, Austin, TX, February 2007.

[119] Abhinav Gupta, Trista Chen, Francine Chen, Don Kimber, and Larry S. Davis. Context and observation driven latent variable model for human pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[120] Abhinav Gupta and Larry S. Davis. Objects in action: An approach for combining action understanding and object perception. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[121] Abhinav Gupta, Anurag Mittal, and Larry S. Davis. Constraint integration for efficient multiview pose estimation with self-occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(3):493–506, March 2008.

[122] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR)*, 3:1157–1182, March 2003.

[123] Lei Han, Wei Liang, Xinxiao Wu, and Yunde Jia. Human action recognition using discriminative models in the learned hierarchical manifold space. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08)*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[124] Tony X. Han, Huazhong Ning, and Thomas S. Huang. Efficient nonparametric belief propagation with application to articulated body tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 214–221,

New York, NY, June 2006.

[125] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W^4s: A real-time system detecting and tracking people in 2 1/2D. In *Proceedings of the European Conference on Computer Vision (ECCV'98) - volume 1*, number 1406 in Lecture Notes in Computer Science, pages 877–892, Freiburg, Germany, June 1998.

[126] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, pages 147–151, Manchester, United Kingdom, August 1988.

[127] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *Annals of Statistics*, 26(2):451–471, April 1998.

[128] Kardelen Hatun and Pınar Duygulu. Pose sentences: A new representation for action recognition using sequence of pose words. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[129] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, February 1983.

[130] Radu Horaud, Matti Niskanen, Guillaume Dewaele, and Edmond Boyer. Human motion tracking by registering an articulated surface to 3-D points and normals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(1):158–163, January 2009.

[131] Shaobo Hou, Aphrodite Galata, Fabrice Caillette, Neil Thacker, and Paul Bromiley. Real-time body tracking using a Gaussian process latent variable model. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[132] Nicholas R. Howe. Silhouette lookup for automatic pose tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)*, page 15, Los Alamitos, CA, June 2004.

[133] Nicholas R. Howe. Flow lookup and biological motion perception. In *Proceedings of the Internation Conference on Image Processing (ICIP'05) - volume 3*, pages 1168–1171, Genova, Italy, September 2005.

[134] Nicholas R. Howe. Boundary fragment matching and articulated pose under occlusion. In *Proceedings of the International Conference on Articulated Motion and Deformable Objects (AMDO'06)*, number 4069 in Lecture Notes in Computer Science, pages 271–280, Port d'Andratx, Spain, July 2006.

[135] Nicholas R. Howe. Recognition-based motion capture and the HumanEva II test data. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (CVPR-EHuM)*, Minneapolis, MN, June 2007.

[136] Nicholas R. Howe. Silhouette lookup for monocular 3D pose tracking. *Image and Vision Computing*, 25(3):331–341, March 2007.

[137] Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. In *Advances in Neural Information Processing Systems (NIPS) 12*, pages 820–826, Denver, CO, November 2000.

[138] Gang Hua and Ying Wu. A decentralized probabilistic approach to articulated body tracking. *Computer Vision and Image Understanding (CVIU)*, 108(3):272–283, December 2007.

[139] Feiyue Huang and Guangyou Xu. Viewpoint insensitive action recognition using en-

velop shape. In *Proceedings of the Asian Conference on Computer Vision (ACCV'07) - part 2*, number 4844 in Lecture Notes in Computer Science, pages 477–486, Tokyo, Japan, November 2007.

[140] Yu Huang and Thomas S. Huang. Model-based human body tracking. In *Proceedings of the International Conference on Pattern Recognition (ICPR'02) - volume 1*, pages 552–555, Quebec, Canada, August 2002.

[141] Zsolt L. Husz, Andrew M. Wallace, and Patrick R. Green. Evaluation of a hierarchical partitioned particle filter with action primitives. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (CVPR-EHuM)*, Minneapolis, MN, June 2007.

[142] Nazlı İkizler, Ramazan G. Cinbiş, and Pınar Duygulu. Human action recognition with line and flow histograms. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[143] Nazlı İkizler, Ramazan G. Cinbiş, Selen Pehlivan, and Pınar Duygulu. Recognizing actions from still images. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[144] Nazlı İkizler and Pınar Duygulu. Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image and Vision Computing*, to appear, 2009.

[145] Nazlı İkizler and David A. Forsyth. Searching for complex human activities with no visual examples. *International Journal of Computer Vision (IJCV)*, 30(3):337–357, December 2008.

[146] Sergey Ioffe and David A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, June 2001.

[147] Michael Isard and Andrew Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, August 1998.

[148] Tomas Izo and Eric L. Grimson. Simultaneous pose recovery and camera registration from multiple views of a walking person. *Image and Vision Computing*, 25(3):342–351, March 2007.

[149] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, March 1991.

[150] Tobias Jaeggli, Esther Koller-Meier, and Luc J. van Gool. Learning generative models for monocular body pose estimation. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[151] Prateek Jain, Brian Kulis, and Kristen Grauman. Fast image search for learned metrics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

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

[153] Kui Jia and Dit-Yan Yeung. Human action recognition using local spatio-temporal discriminant embedding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[154] Hao Jiang and David R. Martin. Finding actions using shape flows. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 2*, number 5303 in Lecture Notes in Computer Science, pages 278–292, Marseille, France, October 2008.

[155] Hao Jiang and David R. Martin. Global pose estimation using non-tree models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

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

[157] Nebojsa Jojic, Jin Gu, Helen Shen, and Thomas S. Huang. 3-D reconstruction of multipart, self-occluding objects. In *Proceedings of the Asian Conference on Computer Vision (ACCV'98)*, pages 455–462, Hong Kong, China, January 1998.

[158] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, second edition, October 2002.

[159] Shanon X. Ju, Michael J. Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'96)*, pages 38–44, Killington, VT, October 1996.

[160] Imran Junejo, Emilie Dexter, Ivan Laptev, and Patrick Pérez. Cross-view action recognition from temporal self-similarities. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 2*, number 5303 in Lecture Notes in Computer Science, pages 293–306, Marseille, France, October 2008.

[161] Timor Kadir and Michael Brady. Scale saliency: A novel approach to salient feature and scale selection. In *proceedings of the International Conference on Visual Information Engineering (VIE)*, pages 25–28, Guildford, United Kingdom, July 2003.

[162] Ioannis A. Kakadiaris and Dimitris N. Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):191–218, December 1998.

[163] Ioannis A. Kakadiaris and Dimitris N. Metaxas. Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12):1453–1459, December 2000.

[164] Atul Kanaujia, Cristian Sminchisescu, and Dimitris Metaxas. Semi-supervised hierarchical models for 3D human pose reconstruction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[165] Atul Kanaujia, Cristian Sminchisescu, and Dimitris Metaxas. Spectral latent variable models for perceptual inference. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[166] Leonard Kaufman and Peter J. Rousseeuw. Clustering by means of medoids. In *Proceedings of the International Conference on Statistical Data Analysis Based on the L1 Norm*, pages 405–416, Neuchâtel, Switzerland, August 1987.

[167] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 506–513, Washington, DC, June 2004.

[168] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 166–173, Beijing, China, October 2005.

[169] Yan Ke, Rahul Sukthankar, and Martial Hebert. Event detection in crowded videos. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[170] Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Proceedings of the Workshop on Visual Surveillance (VS'07)*, pages 1–8, Minneapolis, MN, June 2007.

[171] Roland Kehl and Luc J. van Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):190–209, November 2006.

[172] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Human activity recognition using a dynamic texture based method. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 885–894, Leeds, United Kingdom, September 2008.

[173] Tae-Kyun Kim and Roberto Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear, 2009.

[174] Oliver D. King and David A. Forsyth. How does CONDENSATION behave with a finite number of samples? In *Proceedings of the European Conference on Computer Vision (ECCV'00) - volume 1*, number 1842 in Lecture Notes in Computer Science, pages 695–709, Dublin, Ireland, June 2000.

[175] Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 995–1004, Leeds, United Kingdom, September 2008.

[176] Reinhard Klette and Garry Tee. *Human Motion: Understanding, Modelling, Capture and Animation*, volume 36 of *Computational Imaging and Vision Series*, chapter Understanding Human Motion: A Historic Review, pages 1–22. Springer-Verlag, October 2007.

[177] Ron Kohavi and George H. John. Wrappers for feature selection. *Artificial Intelligence*, 97(1):273–324, December 1997.

[178] Pushmeet Kohli, Jonathan Rihan, Matthieu Bray, and Philip H.S. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision (IJCV)*, 79(3):285–298, September 2009.

[179] Stephen J. Krotosky and Mohan M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Transactions On Intelligent Transportation Systems*, 8(4):619–629, December 2007.

[180] Volker Krüger, Danica Kragic, Aleš Ude, and Christopher Geib. The meaning of action: a review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, September 2007.

[181] Paul Kuo, Thibault Ammar, Michal Lewandowski, Dimitrios Makris, and Jean-Christophe Nebel. Exploiting human bipedal motion constraints for 3D pose recovery from a single uncalibrated camera. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP'09)*, page to appear, Lisboa, Portugal, February 2009.

[182] Paul Kuo, Dimitrios Makris, Najla Megherbi, and Jean-Christophe Nebel. Integration of local image cues for probabilistic 2D pose recovery. In *Proceedings of the International Symposium on Advances in Visual Computing (ISVC'08) - part 2*, number 5359 in Lecture Notes in Computer Science, pages 214–223, Las Vegas, NV, December 2008.

[183] John D. Lafferty, Andrew McCallum, and Fernando C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML'01)*, pages 282–289,

Williamstown, MA, June 2001.

[184] Xiangyang Lan and Daniel P. Huttenlocher. Beyond trees: common-factor models for 2D human pose recovery. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 470–477, Beijing, China, October 2005.

[185] Ivan Laptev, Barbara Caputo, Christian Schüldt, and Tony Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding (CVIU)*, 108(3):207–229, December 2007.

[186] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 1*, pages 432–439, Nice, France, October 2003.

[187] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[188] Ivan Laptev and Patrick Pérez. Retrieving actions in movies. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[189] Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, December 2005.

[190] Chan-Su Lee and Ahmed M. Elgammal. Simultaneous inference of view and body pose using torus manifolds. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06) - volume 3*, pages 489–494, Kowloon Tong, Hong Kong, August 2006.

[191] Hsi-Jian J. Lee and Zen Chen. Determination of 3D human body posture from a single view. *Computer Vision, Graphics and Image Processing*, 30(2):148–168, May 1985.

[192] Mun Wai Lee and Isaac Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 334–341, Washington, DC, June 2004.

[193] Mun Wai Lee and Ram Nevatia. Human pose tracking in monocular sequence using multi-level structured models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(1):27–38, January 2009.

[194] Kobi Levi and Yair Weiss. Learning object detection from a small number of examples: the importance of good features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 53–60, Washington, DC, June 2004.

[195] Rui Li, Tai-Peng Tian, and Stan Sclaroff. Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[196] Rui Li, Ming-Hsuan Yang, Stan Sclaroff, and Tai-Peng Tian. Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers. In *Proceedings of the European Conference on Computer Vision (ECCV'06) - volume 2*, number 3952 in Lecture Notes in Computer Science, pages 137–150, Graz, Austria, May 2006.

[197] David Liebowitz and Stefan Carlsson. Uncalibrated motion capture exploiting articulated structure constraints. *International Journal of Computer Vision*, 51(3):171–187, March 2003.

[198] Zhe Lin and Larry S. Davis. A pose-invariant descriptor for human detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4*, number 5305 in Lecture Notes in Computer Science, pages 423–436, Marseille, France, October 2008.

[199] Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[200] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[201] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[202] Xiaoming Liu, Ting Yu, Thomas Sebastian, and Peter Tu. Boosted deformable model for human body alignment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

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

[204] Wei-Lwun Lu and James J. Little. Simultaneous tracking and action recognition using the PCA-HOG descriptor. In *Proceedings of the Canadian Conference on Computer and Robot Vision (CRV'06)*, pages 6–6, Quebec City, Canada, June 2006.

[205] Yifan Lu. Markerless human motion capture: An application of simulated annealing and fast marching method. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[206] Fengjun Lv and Ram Nevatia. Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost. In *Proceedings of the European Conference on Computer Vision (ECCV'06) - volume 4*, number 3953 in Lecture Notes in Computer Science, pages 359–372, Graz, Austria, May 2006.

[207] Fengjun Lv and Ram Nevatia. Single view human action recognition using key pose matching and Viterbi path searching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[208] John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proceedings of the European Conference on Computer Vision (ECCV'00) - volume 2*, number 1843 in Lecture Notes in Computer Science, pages 3–19, Dublin, Ireland, June 2000.

[209] Étienne-Jules Marey. *La Machine Animale, Locomotion Terrestre et Aérienne*. Germer Baillière, Paris, 1873.

[210] Osama Masoud and Nikos Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, August 2003.

[211] Pyry Matikainen, Martial Hebert, Rahul Sukthankar, and Yan Ke. Fast motion consistency through matrix quantization. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 1055–1064, Leeds, United Kingdom, September 2008.

[212] M. Ángeles Mendoza and Nicolás Pérez de la Blanca. Applying space state models in human action recognition: A comparative study. In *International Workshop on Articulated Motion and Deformable Objects (AMDO'08)*, number 5098 in Lecture Notes in

Computer Science, pages 53–62, Port d'Andratx, Spain, July 2008.

[213] Clement Menier, Edmond Boyer, and Bruno Raffin. 3D skeleton-based body pose recovery. In *Proceedings of the International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 389–396, Chapel Hill, NC, June 2006.

[214] Antonio S. Micilotta, Eng-Jon Ong, and Richard Bowden. Real-time upper body detection and 3D pose estimation in monoscopic images. In *Proceedings of the European Conference on Computer Vision (ECCV'06) - volume 3*, number 3953 in Lecture Notes in Computer Science, pages 139–150, Graz, Austria, May 2006.

[215] Ivana Mikić, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223, July 2003.

[216] Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)*, 60(1):63–86, October 2004.

[217] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(10):1615–1630, Ocotber 2005.

[218] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the European Conference on Computer Vision (ECCV'04) - volume 1*, number 3021 in Lecture Notes in Computer Science, pages 69–82, Prague, Czech Republic, May 2004.

[219] Krystian Mikolajczyk and Hirofumi Uemura. Action recognition with motion-appearance vocabulary forest. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[220] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions On Systems, Man, And Cybernetics (SMC) - Part C: Applications And Reviews*, 37(3):311–324, May 2007.

[221] Anurag Mittal, Liang Zhao, and Larry S. Davis. Human body pose estimation using silhouette shape analysis. In *Proceedings of the Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, pages 263–270, Miami, FL, July 2003.

[222] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):90–126, November 2006.

[223] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(4):349–361, April 2001.

[224] Darnell J. Moore, Irfan A. Essa, and Monson H. Hayes III. Exploiting human actions and object context for recognition tasks. In *Proceedings of the International Conference on Computer Vision (ICCV'99) - volume 1*, pages 80–86, Kerkyra, Greece, September 1999.

[225] Greg Mori and Jitendra Malik. Recovering 3D human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(7):1052–1062, July 2006.

[226] Greg Mori, Xiaofeng Ren, Alexei A. Efros, and Jitendra Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 326–333, Washington, DC, June 2004.

[227] Daniel D. Morris and James M. Rehg. Singularity analysis for articulated object tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pages 289–297, Santa Barbara, CA, June 1998.

[228] Johannes Müller-Gerking, Gert Pfurtscheller, and Henrik Flyvbjerg. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, 110(5):787–798, May 1999.

[229] Lars Mündermann, Stefano Corazza, and Thomas. P. Andriacchi. Markerless human motion capture through visual hull and articulated ICP. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM)*, Whistler, Canada, December 2006.

[230] Lars Mündermann, Stefano Corazza, and Thomas P. Andriacchi. Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[231] Eadweard J. Muybridge. *Animals in Motion*. University of Pennsylvania Press, Philadelphia, PA, 1887.

[232] Pradeep Natarajan and Ram Nevatia. Online, real-time tracking and recognition of human actions. In *Proceedings of the Workshop on Motion and Video Computing (WMVC'08)*, pages 1–8, Copper Mountain, CO, January 2008.

[233] Pradeep Natarajan and Ram Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[234] Ramanan Navaratnam, Andrew W. Fitzgibbon, and Roberto Cipolla. Semi-supervised learning of joint density models for human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC'06) - volume 2*, pages 679–688, Edinburgh, United Kingdom, September 2006.

[235] Ramanan Navaratnam, Andrew W. Fitzgibbon, and Roberto Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[236] Ramanan Navaratnam, Arasanathan Thayananthan, Philip H. Torr, and Roberto Cipolla. Hierarchical part-based human body pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC'05)*, Oxford, United Kingdom, September 2005.

[237] Bingbing Ni, Ashraf Ali Kassim, and Stefan Winkler. A hybrid framework for 3-D human motion tracking. *IEEE Transactions On Circuits And Systems For Video Technology*, 18(8):1075–1084, August 2008.

[238] Juan Carlos Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[239] Juan Carlos Niebles, Bohyung Han, Andras Ferencz, and Li Fei-Fei. Extracting moving people from internet videos. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4*, number 5305 in Lecture Notes in Computer Science, page 527540, Marseille, France, October 2008.

[240] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer*

Vision (IJCV), 79(3):299–318, September 2008.

[241] Huazhong Ning, Yuxiao Hu, and Thomas S. Huang. Searching human behaviors using spatial-temporal words. In *Proceedings of the International Conference on Image Processing (ICIP'07) - volume 6*, pages 337–340, San Antonio, TX, September 2007.

[242] Huazhong Ning, Yuxiao Hu, and Thomas S. Huang. Efficient initialization of mixtures of experts for human pose estimation. In *Proceedings of the International Conference on Image Processing (ICIP'08)*, pages 2164–2167, San Diego, CA, October 2008.

[243] Huazhong Ning, Tieniu Tan, Liang Wang, and Weiming Hu. People tracking based on motion model and motion constraints with automatic initialization. *Pattern Recognition*, 37(7):1423–1440, July 2004.

[244] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas S. Huang. Discriminative learning of visual words for 3D human pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[245] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas S. Huang. Latent pose estimator for continuous action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 2*, number 5305 in Lecture Notes in Computer Science, pages 419–433, Marseille, France, October 2008.

[246] Feng Niu and Mohamed Abdel-Mottaleb. View-invariant human activity recognition based on shape and motion features. In *Proceedings of the International Symposium on Multimedia Software Engineering (ISMSE'04)*, pages 546–556, Miami, FL, December 2004.

[247] Sourabh A. Niyogi and Edward H. Adelson. Analyzing and recognizing walking figures in XYT. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 469–474, Seattle, WA, June 1994.

[248] Sebastian Nowozin, Gökhan Bakır, and Koji Tsuda. Discriminative subsequence mining for action classification. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[249] Abhijit S. Ogale, Alap Karapurkar, and Yiannis Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *Revised Papers of the Workshops on Dynamical Vision (WDV'05 and WDV'06)*, number 4358 in Lecture Notes in Computer Science, pages 115–126, Beijing, China, May 2007.

[250] Takehito Ogata, William Christmas, Josef Kittler, and Seiji Ishikawa. Improving human activity detection by combining multi-dimensional motion descriptors with boosting. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06) - volume 1*, pages 295–298, Kowloon Tong, Hong Kong, August 2006.

[251] Antonios Oikonomopoulos, Maja Pantic, and Ioannis Patras. B-spline polynomial descriptors for human activity recognition. In *Proceedings of the Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB'08)*, pages 1–8, Anchorage, AK, June 2008.

[252] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions On Systems, Man, And Cybernetics (SMC) - Part B: Cybernetics*, 36(3):710–719, June 2006.

[253] Ryuzo Okada and Stefano Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 2*, number 5303 in Lecture Notes in Computer

Science, pages 434–445, Marseille, France, October 2008.

[254] Eng-Jon Ong and Shaogang Gong. A dynamic 3D human model using hybrid 2D-3D representations in hierarchical pca space. In *Proceedings of the British Machine Vision Conference (BMVC'99)*, pages 33–42, Nottingham, United Kingdom, September 1999.

[255] Eng-Jon Ong, Antonio S. Micilotta, Richard Bowden, and Adrian Hilton. Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):178–189, November 2006.

[256] Katsunori Onishi, Tetsuya Takiguchi, and Yasuo Ariki. 3D human posture estimation using the HOG features from monocular image. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[257] Joseph O'Rourke and Norman I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2(6):522–536, November 1980.

[258] Carlos Orrite-Uruñuela, Francisco Martínez, José E. Herrero-Jaraba, Hossein Ragheb, and Sergio Velastin. Independent viewpoint silhouette-based human action modelling and recognition. In *Proceedings of the International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA'08)*, pages 1–12, Marseille, France, October 2008.

[259] Olusegun Oshin, Andrew Gilbert, John Illingworth, and Richard Bowden. Spatio-temporal feature recognition using randomised ferns. In *Proceedings of the International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA'08)*, pages 1–12, Marseille, France, October 2008.

[260] Junbiao Pang, Laiyun Qing, Qingming Huang, Shuqiang Jiang, and Wen Gao. Monocular tracking 3D people by gaussian process spatio-temporal variable model. In *Proceedings of the International Conference on Image Processing (ICIP'07) - volume 5*, pages 41–44, San Antonio, TX, September 2007.

[261] Vasu Parameswaran and Rama Chellappa. View independent human body pose estimation from a single perspective image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 2*, pages 16–22, Washington, DC, June 2004.

[262] Vasu Parameswaran and Rama Chellappa. View invariance for human action recognition. *International Journal of Computer Vision (IJCV)*, 66(1):83–101, January 2006.

[263] Sangho Park and Mohan M. Trivedi. Understanding human interactions with track and body synergies (TBS) captured from multiple views. *Computer Vision and Image Understanding (CVIU)*, 111(1):2–20, July 2008.

[264] Alonso Patron-Perez and Ian Reid. A probabilistic framework for recognizing similar actions using spatio-temporal features. In *Proceedings of the British Machine Vision Conference (BMVC'07)*, pages 1–10, Edinburgh, United Kingdom, September 2007.

[265] Vladimir I. Pavlović, James M. Rehg, Tat-Jen Cham, and Kevin P. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proceedings of the International Conference on Computer Vision (ICCV'99) - volume 1*, pages 94–101, Kerkyra, Greece, September 1999.

[266] Vladimir I. Pavlović, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):677–695, July 1997.

[267] Patrick Peursum, Svetha Venkatesh, and Geoff West. Observation-switching linear dy-

namic systems for tracking humans through unexpected partial occlusions by scene objects. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06) - volume 4*, pages 929–934, Kowloon Tong, Hong Kong, August 2006.

[268] Patrick Peursum, Svetha Venkatesh, and Geoff West. Tracking-as-recognition for articulated full-body human motion analysis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[269] Patrick Peursum, Svetha Venkatesh, and Geoff West. A study on smoothing for particle-filtered 3D human body tracking. *International Journal of Computer Vision (IJCV)*, to appear, 2009.

[270] Rolf Plänkers and Pascal Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding (CVIU)*, 81(3):285–302, March 2001.

[271] Ramprasad Polana and Randal C. Nelson. Detection and recognition of periodic, non-rigid motion. *International Journal of Computer Vision (IJCV)*, 23(3):261–282, June 1997.

[272] Ronald Poppe. Evaluating example-based pose estimation: Experiments on the HumanEva sets. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (CVPR-EHuM)*, Minneapolis, MN, June 2007.

[273] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):4–18, October 2007.

[274] Ronald Poppe and Mannes Poel. Example-based pose estimation in monocular images using compact Fourier descriptors. CTIT Technical report 05-49, University of Twente, Enschede, The Netherlands, October 2005.

[275] Ronald Poppe and Mannes Poel. Comparison of silhouette shape descriptors for example-based human pose recovery. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'06)*, pages 541–546, Southampton, United Kingdom, April 2006.

[276] Ronald Poppe and Mannes Poel. Body-part templates for recovery of 2D human poses under occlusion. In *International Workshop on Articulated Motion and Deformable Objects (AMDO'08)*, number 5098 in Lecture Notes in Computer Science, pages 289–298, Port d'Andratx, Spain, July 2008.

[277] Ronald Poppe and Mannes Poel. Discriminative human action recognition using pairwise CSP classifiers. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08)*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[278] Fatih Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1*, pages 829–836, San Diego, CA, June 2005.

[279] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. Human action recognition using robust power spectrum features. In *Proceedings of the International Conference on Image Processing (ICIP'08)*, pages 753–756, San Diego, CA, October 2008.

[280] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. ViHASi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In *Proceedings of the International Conference on Distributed Smart Cameras (ICDSC'08)*, pages 1–10, Stanford, CA, September 2008.

[281] Deva Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems (NIPS) 19*, pages 1129–1136, Vancouver, Canada,

December 2006.

[282] Deva Ramanan and David A. Forsyth. Automatic annotation of everyday movements. In *Advances in Neural Information Processing Systems (NIPS) 16*, pages 1–8, Vancouver, Canada, December 2003.

[283] Deva Ramanan, David A. Forsyth, and Andrew Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(1):65–81, January 2007.

[284] Deva Ramanan and Cristian Sminchisescu. Training deformable models for localization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 206–213, New York, NY, June 2006.

[285] Cen Rao, Alper Yilmaz, and Mubarak Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision (IJCV)*, 50(2):203–226, November 2002.

[286] Konstantinos Rapantzikos, Yannis S. Avrithis, and Stefanos D. Kollias. Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: Potential in human action recognition. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR'07)*, pages 294–301, Amsterdam, The Netherlands, July 2007.

[287] Liu Ren, Gregory Shakhnarovich, Jessica K. Hodgins, Hanspeter Pfister, and Paul A. Viola. Learning silhouette features for control of human motion. *ACM Transactions on Computer Graphics*, 24(4):1303–1331, October 2005.

[288] Xiaofeng Ren, Alexander C. Berg, and Jitendra Malik. Recovering human body configurations using pairwise constraints between parts. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 824–831, Beijing, China, October 2005.

[289] Timothy J. Roberts, Stephen J. McKenna, and Ian W. Ricketts. Human tracking using 3D surface colour distributions. *Image and Vision Computing*, 24(12):1332–1342, December 2006.

[290] Neil Robertson and Ian Reid. A general method for human activity recognition in video. *Computer Vision and Image Understanding (CVIU)*, 104(2):232–248, November 2006.

[291] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[292] Grégory Rogez, José J. Guerrero, Jesús Martínez, and Carlos Orrite-Uruñuela. Viewpoint independent human motion analysis in man-made environments. In *Proceedings of the British Machine Vision Conference (BMVC'06) - volume 2*, pages 659–668, Edinburgh, United Kingdom, September 2006.

[293] Grégory Rogez, Carlos Orrite-Uruñuela, and Jesús Martínez del Rincón. A spatio-temporal 2D-models framework for human pose recovery in monocular sequences. *Pattern Recognition*, 41(9):2926–2944, September 2008.

[294] Grégory Rogez, Jonathan Rihan, Srikumar Ramalingam, Carlos Orrite-Uruñuela, and Philip H.S. Torr. Randomized trees for human pose detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[295] Karl Rohr. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 59(1):94–115,

January 1994.

[296] Rémi Ronfard, Cordelia Schmid, and Bill Triggs. Learning to parse pictures of people. In *Proceedings of the European Conference on Computer Vision (ECCV'02) - volume 4*, number 2353 in Lecture Notes in Computer Science, pages 700–714, Copenhagen, Denmark, May 2002.

[297] Rómer E. Rosales. Recognition of human action using moment-based features. Technical Report BU-1998-020, Boston University, Computer Science, Boston, MA, November 1998.

[298] Rómer E. Rosales and Stan Sclaroff. Learning body pose via specialized maps. In *Advances in Neural Information Processing Systems (NIPS) 14*, pages 1263–1270, Vancouver, Canada, December 2001.

[299] Rómer E. Rosales and Stan Sclaroff. Combining generative and discriminative models in a framework for articulated pose estimation. *International Journal of Computer Vision (IJCV)*, 67(3):251–276, May 2006.

[300] Rómer E. Rosales, Matheen Siddiqui, Jonathan Alon, and Stan Sclaroff. Estimating 3D body pose using uncalibrated cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01) - volume 1*, pages 821–827, Kauai, HI, December 2001.

[301] Bodo Rosenhahn, Uwe Kersting, Katie Powell, Reinhard Klette, Gisela Klette, and Hans-Peter Seidel. A system for articulated tracking incorporating a clothing model. *Machine Vision and Applications*, 18(1):25–40, February 2007.

[302] Bodo Rosenhahn, Christian Schmaltz, Thomas Brox, Joachim Weickert, Daniel Cremers, and Hans-Peter Seidel. Markerless motion capture of man-machine interaction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[303] Bodo Rosenhahn, Christian Schmaltz, Thomas Brox, Joachim Weickert, and Hans-Peter Seidel. Staying well grounded in markerless motion capture. In *Proceedings of the Symposium on Pattern Recognition (DAGM'08)*, number 5096 in Lecture Notes in Computer Science, pages 385–395, Munich, Germany, June 2008.

[304] Michael S. Ryoo and Jake K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision (IJCV)*, 82(1):1–24, April 2009.

[305] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, and Li Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *Proceedings of the Workshop on Applications of Computer Vision (WACV'08)*, pages 1–8, Copper Mountain, CO, January 2008.

[306] Konrad Schindler and Luc J. van Gool. Action snippets: How many frames does human action recognition require? In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[307] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the International Conference on Pattern Recognition (ICPR'04) - volume 3*, pages 32–36, Cambridge, United Kingdom, August 2004.

[308] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *Proceedings of the International Conference on Multimedia (MultiMedia'07)*, pages 357–360, Augsburg, Germany, September 2007.

[309] Steven M. Seitz and Charles R. Dyer. View-invariant analysis of cyclic motion. *Interna-*

tional Journal of Computer Vision (IJCV), 25(3):231–251, December 1997.

[310] Gregory Shakhnarovich, Paul A. Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 2*, pages 750–759, Nice, France, October 2003.

[311] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[312] Eli Shechtman and Michal Irani. Space-time behavior-based correlation — OR — How to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):2045–2056, November 2007.

[313] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of a human action. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 144–149, Beijing, China, October 2005.

[314] Yuping Shen and Hassan Foroosh. View-invariant recognition of body pose from space-time templates. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–6, Anchorage, AK, June 2008.

[315] Qinfeng Shi, Li Wang, Li Cheng, and Alex Smola. Discriminative human action segmentation and recognition using semi-Markov model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[316] Matheen Siddiqui and Gérard Medioni. Efficient upper body pose estimation from a single image or a sequence. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 74–87, Rio de Janeiro, Brazil, October 2007.

[317] Hedvig Sidenbladh and Michael J. Black. Learning the statistics of people in images and video. *International Journal of Compututer Vision*, 54(1–3):181–207, August 2003.

[318] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proceedings of the European Conference on Computer Vision (ECCV'00) - volume 2*, number 1843 in Lecture Notes in Computer Science, pages 702–718, Dublin, Ireland, June 2000.

[319] Hedvig Sidenbladh, Michael J. Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV'02) - volume 1*, number 2350 in Lecture Notes in Computer Science, pages 784–800, Copenhagen, Denmark, May 2002.

[320] Leonid Sigal, Alexandru O. Bălan, and Michael J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems (NIPS) 20*, pages 1337–1344, Vancouver, Canada, December 2008.

[321] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. Tracking loose-limbed people. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04) - volume 1*, pages 421–428, Washington, DC, June 2004.

[322] Leonid Sigal and Michael J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI, September 2006.

[323] Leonid Sigal and Michael J. Black. Measure locally, reason globally: Occlusion-sensitive

articulated pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 2041–2048, New York, NY, June 2006.

[324] Leonid Sigal and Michael J. Black. Predicting 3D people from 2D pictures. In *Proceedings of the International Conference on Articulated Motion and Deformable Objects (AMDO'06)*, number 4069 in Lecture Notes in Computer Science, pages 185–195, Port d'Andratx, Spain, July 2006.

[325] Leonid Sigal, Michael Isard, Benjamin Sigelman, and Michael J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Advances in Neural Information Processing Systems (NIPS) 16*, pages 1539–1546, Vancouver, Canada, December 2003.

[326] Cristian Sminchisescu and Allan D. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *Proceedings of the International Conference on Machine Learning (ICML'04)*, pages 759–766, Banff, Canada, July 2004.

[327] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Learning joint top-down and bottom-up processes for 3D visual inference. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1743–1752, New York, NY, June 2006.

[328] Cristian Sminchisescu, Atul Kanaujia, and Dimitris N. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):210–220, November 2006.

[329] Cristian Sminchisescu, Atul Kanaujia, and Dimitris N. Metaxas. $\mathrm{BM}^3\mathrm{E}$: Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):2030–2044, November 2007.

[330] Cristian Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotic Research*, 22(6):371–392, June 2003.

[331] Cristian Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03) - volume 1*, pages 69–76, Madison, WI, June 2003.

[332] Paul Smith, Niels da Vitoria Lobo, and Mubarak Shah. TemporalBoost for event recognition. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 733–740, Beijing, China, October 2005.

[333] Yang Song, Luis Goncalves, and Pietro Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(7):814–827, July 2003.

[334] Richard Souvenir and Justin Babbs. Learning the viewpoint manifold for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–7, Anchorage, AK, June 2008.

[335] Jonathan Starck and Adrian Hilton. Model-based human shape reconstruction from multiple views. *Computer Vision and Image Understanding (CVIU)*, 111(2):179–194, August 2008.

[336] Josephine Sullivan and Stefan Carlsson. Recognizing and tracking human action. In *Proceedings of the European Conference on Computer Vision (ECCV'02) - volume 1*, number 2350 in Lecture Notes in Computer Science, pages 629–644, Copenhagen, Denmark, May 2002.

[337] Evan A. Suma, Christopher W. Sinclair, Justin Babbs, and Richard Souvenir. A sketch-

based approach for detecting common human actions. In *Proceedings of the International Symposium on Advances in Visual Computing (ISVC'08) - part 1*, number 5358 in Lecture Notes in Computer Science, pages 418–427, Las Vegas, NV, December 2008.

[338] Yunda Sun, Matthieu Bray, Arasanathan Thayananthan, Baozong Yuan, and Philip H.S. Torr. Regression-based human motion capture from voxel data. In *Proceedings of the British Machine Vision Conference (BMVC'06) - volume 1*, pages 277–286, Edinburgh, United Kingdom, September 2006.

[339] Aravind Sundaresan and Rama Chellappa. Model driven segmentation of articulating humans in laplacian eigenspace. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(10):1771–1785, October 2008.

[340] Therdsak Tangkuampien and David Suter. Real-time human pose inference using kernel principal component pre-image approximations. In *Proceedings of the British Machine Vision Conference (BMVC'06) - volume 2*, pages 599–608, Edinburgh, United Kingdom, September 2006.

[341] Leonid Taycher, Gregory Shakhnarovich, David Demirdjian, and Trevor Darrell. Conditional random people: Tracking humans with crfs and grid filters. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 222–229, New York, NY, June 2006.

[342] Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding (CVIU)*, 80(3):349–363, December 2000.

[343] Yee W. Teh and Sam T. Roweis. Automatic alignment of local representations. In *Advances in Neural Information Processing Systems (NIPS) 15*, pages 841–848, Vancouver, Canada, December 2002.

[344] Arasanathan Thayananthan, Ramanan Navaratnam, Bjoörn Stenger, Philip H.S. Torr, and Roberto Cipolla. Pose estimation and tracking using multivariate regression. *Pattern Recognition Letters*, 29(9):1302–1310, July 2003.

[345] Christian Thurau. Behavior histograms for action recognition and human detection. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil, October 2007.

[346] Christian Thurau and Václav Hlaváč. Pose primitive based human action recognition in videos or still images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–6, Anchorage, AK, June 2008.

[347] Tai-Peng Tian, Rui Li, and Stan Sclaroff. Tracking human body pose on a learned smooth space. Technical Report BUCS-TR-2005-029, Boston University, Computer Science Department, Boston, MA, July 2005.

[348] Kentaro Toyama and Andrew Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1):9–19, June 2002.

[349] Du Tran, Alexander Sorokin, and David A. Forsyth. Human activity recognition with metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 1*, number 5302 in Lecture Notes in Computer Science, pages 548–561, Marseille, France, October 2008.

[350] Roger Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.

[351] Pavan K. Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions On Circuits And Systems For Video Technology*, 18(11):1473–1488, November 2008.

[352] Pavan K. Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[353] Pavan K. Turaga, Ashok Veeraraghavan, and Rama Chellappa. Unsupervised view and rate invariant clustering of video sequences. *Computer Vision and Image Understanding (CVIU)*, 113(3):353–371, March 2009.

[354] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, January 2008.

[355] Hirofumi Uemura, Seiji Ishikawa, and Krystian Mikolajczyk. Feature tracking and motion compensation for action recognition. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 293–302, Leeds, United Kingdom, September 2008.

[356] Norimichi Ukita, Ryosuke Tsuji, and Masatsugu Kidode. Real-time shape analysis of a human body in clothing using time-series part-labeled volumes. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 3*, number 5304 in Lecture Notes in Computer Science, pages 681–695, Marseille, France, October 2008.

[357] Raquel Urtasun and Trevor Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[358] Raquel Urtasun, David J. Fleet, and Pascal Fua. Monocular 3-D tracking of the golf swing. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 2*, pages 932–938, San Diego, CA, June 2005.

[359] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3D people tracking with Gaussian process dynamical models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 238–245, New York, NY, June 2006.

[360] Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor Darrell, and Neil D. Lawrence. Topologically-constrained latent variable models. In *Proceedings of the International Conference on Machine Learning (ICML'08)*, pages 1080–1087, Helsinki, Finland, July 2008.

[361] Raquel Urtasun, David J. Fleet, Aaron Hertzmann, and Pascal Fua. Priors for people tracking from small training sets. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 1*, pages 403–410, Beijing, China, October 2005.

[362] Raquel Urtasun, David J. Fleet, and Neil D. Lawrence. Modeling human locomotion with topologically constrained latent variable models. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 104–118, Rio de Janeiro, Brazil, October 2007.

[363] Ashok Veeraraghavan, Rama Chellappa, and Amit K. Roy-Chowdhury. The function space of an activity. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 959–968, New York, NY, June 2006.

[364] Ashok Veeraraghavan, Amit K. Roy-Chowdhury, and Rama Chellappa. Matching shape sequences in video with applications in human movement analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(12):1896–1909, December 2005.

[365] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01) - volume 1*, pages 511–518, Kauai, HI, December 2001.

[366] Shiv N. Vitaladevuni, Vili Kellokumpu, and Larry S. Davis. Action recognition using ballistic dynamics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[367] Daniel Vlasic, Ilya Baran, Wojciech Matusiky, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):1–9, August 2008.

[368] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Physical simulation for probabilistic motion tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[369] Stefan Wachter and Hans-Hellmut Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding (CVIU)*, 74(3):174–192, June 1999.

[370] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Multifactor Gaussian process models for style-content separation. In *Proceedings of the International Conference on Machine Learning (ICML'07)*, number 227 in ACM International Conference Proceeding Series, pages 975–982, Corvalis, OR, June 2007.

[371] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):283–298, Februari 2008.

[372] Jessica J. Wang and Sameer Singh. Video analysis of human dynamics: a survey. *Real-Time Imaging*, 9(5):321–346, October 2003.

[373] Liang Wang, Weiming Hu, and Tieniu Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, March 2003.

[374] Liang Wang and David Suter. Informative shape representations for human action recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06) - volume 2*, pages 1266–1269, Kowloon Tong, Hong Kong, August 2006.

[375] Liang Wang and David Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Transactions On Image Processing (TIP)*, 16(6):1646–1661, June 2007.

[376] Liang Wang and David Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[377] Liang Wang and David Suter. Visual learning and recognition of sequential data manifolds with applications to human movement analysis. *Computer Vision and Image Understanding (CVIU)*, 110(2):153–172, May 2008.

[378] Ping Wang and James M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 1*, pages 790–797, New York, NY, June 2006.

[379] Ruixuan Wang, Wee Kheng Leow, and Hon Wai Leong. 3D-2D spatiotemporal registration for sports motion analysis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, AK, June 2008.

[380] Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1654–1661, New York, NY, June 2006.

[381] Yang Wang and Greg Mori. Learning a discriminative hidden part model for human action recognition. In *Advances in Neural Information Processing Systems (NIPS) 21*, pages 1721–1728, Vancouver, Canada, December 2008.

[382] Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 3*, number 5304 in Lecture Notes in Computer Science, pages 710–724, Marseille, France, October 2008.

[383] Yang Wang, Payam Sabzmeydani, and Greg Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 240–254, Rio de Janeiro, Brazil, October 2007.

[384] Ying Wang, Kaiqi Huang, and Tieniu Tan. Human activity recognition based on $\Re$ transform. In *Proceedings of the Workshop on Visual Surveillance (VS'07)*, pages 1–8, Minneapolis, MN, June 2007.

[385] Daniel Weinland and Edmond Boyer. Action recognition using exemplar-based embedding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–7, Anchorage, AK, June 2008.

[386] Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3D exemplars. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[387] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Automatic discovery of action taxonomies from multiple views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1639–1645, New York, NY, June 2006.

[388] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):249–257, November 2006.

[389] D. Randall Wilson and Tony R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, March 2000.

[390] Shu-Fai Wong and Roberto Cipolla. Extracting spatiotemporal interest points using global information. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[391] Shu-Fai Wong, Tae-Kyun Kim, and Roberto Cipolla. Learning motion categories using both semantic and structural information. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–8, Minneapolis, MN, June 2007.

[392] Christopher R. Wren, Ali J. Azarbayejani, Trevor Darrell, and Alex P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):780–785, July 1997.

[393] Bo Wu and Ram Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *International Journal of Computer Vision (IJCV)*, 75(2):247–266, November 2007.

[394] Bo Wu and Ram Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *International Journal*

of Computer Vision (IJCV), 82(2):185–204, April 2009.

[395] Chen Wu and Hamid K. Aghajan. Real-time human pose estimation: A case study in algorithm design for smart camera networks. *Proceedings of the IEEE*, 96(10):1715–1732, October 2008.

[396] Xinxiao Wu, Wei Liang, and Yunde Jia. Tracking articulated objects by learning intrinsic structure of motion. *Pattern Recognition Letters*, 30(3):267274, February 2009.

[397] Li-Qun Xu and David C. Hogg. Neural networks in human motion tracking - an experimental study. *Image and Vision Computing*, 15(8):607–615, August 1997.

[398] Xinyu Xu and Baoxin Li. Learning motion correlation for tracking articulated human body with a rao-blackwellised particle filter. In *Proceedings of the International Conference On Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[399] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'00) - volume 1*, pages 151–156, Hilton Head Island, SC, June 2000.

[400] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'92)*, pages 379–385, Champaign, IL, June 1992.

[401] Pingkun Yan, Saad M. Khan, and Mubarak Shah. Learning 4D action feature models for arbitrary view action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–7, Anchorage, AK, June 2008.

[402] Changjiang Yang, Yanlin Guo, Harpreet S. Sawhney, and Rakesh Kumar. Learning actions using robust string kernels. In *Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07)*, number 4814 in Lecture Notes in Computer Science, pages 313–327, Rio de Janeiro, Brazil, October 2007.

[403] Hee-Deok Yang and Seong-Whan Lee. Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. *Pattern Recognition*, 40(11):3120–3131, November 2007.

[404] Alper Yilmaz and Mubarak Shah. Matching actions in presence of camera motion. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):221–231, November 2006.

[405] Alper Yilmaz and Mubarak Shah. A differential geometric approach to representing the human actions. *Computer Vision and Image Understanding (CVIU)*, 119(3):335–351, March 2008.

[406] Lihi Zelnik-Manor and Michal Irani. Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(9):1530–1535, September 2006.

[407] Xin Zhang and Guoliang Fan. Dual generative models for human motion estimation from an uncalibrated monocular camera. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[408] Ziming Zhang, Yiqun Hu, Syin Chan, and Liang-Tien Chia. Motion context: A new representation for human action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4*, number 5305 in Lecture Notes in Computer Science, pages 817–829, Marseille, France, October 2008.

[409] Tao Zhao and Ram Nevatia. 3D tracking of human locomotion: A tracking as recognition approach. In *Proceedings of the International Conference on Pattern Recognition*

(ICPR'02) - volume 1, pages 546–551, Quebec, Canada, August 2002.

[410] Xu Zhao, Huazhong Ning, Yuncai Liu, and Thomas S. Huang. Discriminative estimation of 3D human pose using gaussian processes. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[411] Zhipeng Zhao and Ahmed Elgammal. Human activity recognition from frame's spatiotemporal representation. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, Tampa, FL, December 2008.

[412] Zhipeng Zhao and Ahmed Elgammal. Information theoretic key frame selection for action recognition. In *Proceedings of the British Machine Vision Conference (BMVC'08)*, pages 1095–1104, Leeds, United Kingdom, September 2008.

[413] Guangyu Zhu, Changsheng Xu, Wen Gao, and Qingming Huang. Action recognition in broadcast tennis video using optical flow and support vector machine. In *Proceedings of the Workshop on Computer Vision in Human-Computer Interaction (ECCV-HCI'06)*, number 3979 in Lecture Notes in Computer Science, pages 89–98, Graz, Austria, May 2007.

[414] Qiang Zhu, Shai Avidan, Mei-Chen Yeh, and Kwang-Ting Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1491–1498, New York, NY, June 2006.

Summary

The automatic analysis of human motion from images and video opens up the way for many applications in the domains of security and surveillance, human-computer interaction, animation, retrieval and sports motion analysis, to name a few. As the analysis of human motion comprises many aspects, in this dissertation, the focus is limited to *human pose recovery* and *human action recognition*. The former is a regression task where the aim is to determine the locations or angles of key joints in the human body, given an image of a human figure. The latter is the process of labeling image sequences with action labels, which is a classification task. The focus is on robust and fast recovery and recognition of human motion. The work is evaluated extensively on publicly available datasets.

An example-based pose recovery approach is introduced where a variant of histograms of oriented gradients (HOG) is used as the image descriptor. It is assumed that a region of interest (ROI) containing a human subject is available, and that the foreground labeling can be determined. Given an unseen image, the ROI is represented as a HOG descriptor. From an example database containing thousands of HOG-pose pairs, the visually closest examples are selected. Weighted interpolation of the corresponding poses is used to obtain the pose estimate, represented as the 3D locations of 20 key joints in the human body. This approach is fast due to the use of a low-cost distance function. Experiments have been carried out on the HumanEva dataset. When using a single camera, mean 3D relative errors of 65 *mm* per joint were obtained, and 45 *mm* per joint on walking and jogging sequences. Combining the input from multiple views slightly decreased the estimation error. In additional experiments, the effect on the accuracy was examined when varying the HOG grid size and the number of examples in the example set.

Partial occlusion of the human figure in the image is a common problem in realistic settings, but has largely been ignored in pose recovery approaches. In this work, the issue is explicitly addressed and it is assumed that a prediction of the occluded areas can be obtained together with a ROI estimate. The example-based approach is adapted to cope with these partial occlusions. The normalization and matching of the HOG descriptors was changed from global to the cell level. The approach was used for recovery of human poses without adjusting the example set and matching process to the occlusion condition. Experiments on the HumanEva dataset with simulated occlusion showed that occlusions affected the recovery accuracy but only moderately.

For the recognition of human actions, again a variant of HOG descriptors is used to encode the ROI of an image containing a human subject. Simple functions are used to discriminate between two classes after applying a common spatial patterns (CSP) transform on sequences of these HOG descriptors. In the transform, the difference in variance between two classes is maximized. Each of the discriminative functions softly votes into the two classes. After evaluation of all pairwise functions, the action class that receives most of the voting mass is the estimated class. Using this approach, approximately 95% was classified correctly on the Weizmann human action dataset and showed that state of the art performance can be obtained

with low training requirements. Additional experiments showed that walking motion could be recognized even for moderate viewpoint changes and with several image deformations. Also, good results were reported when shorter sequences were used.

To deal with different viewpoints and partial occlusion of the human figure in the image, the example-based pose recovery approach is combined with the CSP-based human action recognition approach. The resulting approach was able to simultaneously recover human poses and recognize the action over a sequence of frames. Specifically, the recovered 3D poses were normalized for rotation and body size of the subject and used as input for the CSP classifier. Again, the HumanEva dataset was used for experiments. Thanks to the rotation normalization, actions could be recognized from arbitrary viewpoints. By handling occlusions in the pose recovery step, action could be recognized from image observations where occlusion was simulated. Finally, the potential of this approach for temporal action segmentation was demonstrated.

Samenvatting

Automatische bewegingsanalyse van het menselijk lichaam in foto's en video heeft vele toepassingen, bijvoorbeeld op het gebied van beveiliging en bewaking, mens-machine interactie, animatie, het ontsluiten van informatie en in de sport. Bewegingsanalyse is een breed gebied, waarbij de focus in dit proefschrift gelegd is op het bepalen van lichaamsposes (*human pose recovery*). Hierbij is het doel om een numerieke benadering te vinden van de stand of locatie van de belangrijkste gewrichten in het menselijk lichaam. Daarnaast richt dit werk zich op het classificeren van menselijke acties (*human action recognition*), waarbij een label uit een beperkte set wordt toegekend aan een opeenvolging van beelden over tijd. De focus heeft gelegen op het creëren van robuuste en snelle algoritmes voor het bepalen van poses en het toekennen van actie-labels. De ontwikkelde methodes zijn geëvalueerd met behulp van vrij beschikbare datasets.

In dit proefschrift wordt een methode beschreven voor het bepalen van poses waarbij het beeld wordt gecodeerd als een histogram van georiënteerde gradiënten (HOG). De aanname is dat een beeldregio met daarin een persoon (ROI) vooraf is bepaald, en dat een scheiding van voor- en achtergrond gemaakt kan worden. De ROI van een te analyseren beeld wordt gecodeerd als een HOG-beschrijving. Uit een database met duizenden opgeslagen pose-HOG paren worden vervolgens de meeste gelijkende geselecteerd. De bijbehorende poses, de 3D locaties van 20 gewrichten, worden naar ratio geïnterpoleerd om tot een benadering van de pose te komen. Deze aanpak is snel doordat de gelijkenis van twee HOG-beschrijvingen direct te bepalen is. De HumanEva dataset is gebruikt om deze aanpak te evalueren. Met beelden van een enkele camera gaf deze aanpak gemiddelde fouten van 65 *mm* per gewricht, gemeten relatief ten opzichte van het centrum van het lichaam. Voor loop- en jogbewegingen was deze fout 45 *mm*. Als beelden uit meerdere camera's werden gecombineerd, was de fout iets lager. In verdere experimenten is de invloed op de nauwkeurigheid onderzocht van variaties op de HOG-beschrijving en het aantal voorbeelden in de database.

Gedeeltelijke occlusie van de persoon in het beeld is een veel voorkomend gegeven in realistische situaties, maar er is weinig onderzoek naar gedaan. In dit werk is expliciet gekeken naar deze occlusies waarbij is aangenomen dat, naast een ROI, een schatting beschikbaar is welke delen van de persoon niet zichtbaar zijn. De database-aanpak is als basis gebruikt om met deze partiële observaties om te gaan. De normalisatie en het vergelijken van HOG-beschrijvingen is aangepast door dit deelsgewijs te doen, in plaats van als geheel. Deze methode is gebruikt om poses te bepalen zonder de database te hoeven aanpassen aan de specifieke mate van overlap. Experimenten op de HumanEva dataset met gesimuleerde occlusies toonden aan dat occlusies slechts een beperkt effect op de nauwkeurigheid hebben.

Voor het labelen van acties zijn wederom HOG-beschrijvingen gebruikt om het beeld binnen de ROI te coderen. Na het uitvoeren van *common spatial patterns* (CSP) op een reeks van opeenvolgende beelden, zijn eenvoudige functies gebruikt om onderscheid te maken tussen twee klassen. De CSP transformatie maakt gebruik van het verschil in variantie tussen de twee klassen en maximaliseert deze. Elk van de scheidende functies verdeelt een eenheids-

massa over de twee klassen. Na het evalueren van al deze functies wordt de klasse gekozen die de meeste massa heeft gekregen. Deze aanpak classificeerde 95% correct op de Weizmann dataset. Dit toont aan dat competitieve resultaten behaald kunnen worden zonder een bewerkelijke trainingsfase. Verder konden loopbewegingen worden herkend ondanks afwijkingen in de kijkrichting tot 70°. Ook wist deze methode om te gaan met verschillende variaties in het beeld, en met kortere beeldreeksen.

Om acties te kunnen labelen met grotere afwijkingen in kijkrichting en met gedeeltelijke occlusie, zijn de methoden voor pose benadering en actie classificatie gecombineerd. Met deze gecombineerde aanpak kunnen reeksen beelden worden gelabeld en kunnen tegelijkertijd de poses van de persoon worden geschat. In deze aanpak wordt voor elk beeld eerst de 3D pose benaderd, die vervolgens onafhankelijk gemaakt wordt van de oriëntatie en lengte van de persoon. Deze genormaliseerde pose wordt vervolgens gebruikt als invoer voor de CSP classificator. Wederom zijn experimenten uitgevoerd op de HumanEva dataset. Dankzij de onafhankelijkheid van oriëntatie konden beeldreeksen correct worden gelabeld vanuit willekeurige kijkrichtingen. Loopbewegingen konden worden herkend uit beeldreeksen met gesimuleerde occlusie dankzij het oplossen van de occlusie in de pose benaderingsstap. Tenslotte is aangetoond dat deze methode ook gebruikt kan worden om beeldreeksen op te delen in segmenten waarbij in ieder segment een andere actie uitgevoerd wordt.

SIKS dissertation series

Since 1998, all dissertations written by PhD students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series. This dissertation is the 206^{th} in the series.

2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion*

2009-06 Muhammad Subianto (UU), *Understanding Classification*

2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*

2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT*

2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques*

2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models*

2008-35 Ben Torben Nielsen (UvT), *Dendritic morphologies: function shapes structure*

2008-34 Jeroen de Knijf (UU), *Studies in Frequent Tree Mining*

2008-33 Frank Terpstra (UVA), *Scientific Workflow Design; theoretical and practical issues*

2008-32 Trung H. Bui (UT), *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes*

2008-31 Loes Braun (UM), *Pro-Active Medical Information Retrieval*

2008-30 Wouter van Atteveldt (VU), *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*

2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines – Of Annotators, Embodied Agents, Users, and Other Humans*

2008-28 Ildiko Flesch (RUN), *On the Use of Independence Relations in Bayesian Networks*

2008-27 Hubert Vogten (OU), *Design and Implementation Strategies for IMS Learning Design*

2008-26 Marijn Huijbregts (UT), *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*

2008-25 Geert Jonker (UU), *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*

2008-24 Zharko Aleksovski (VU), *Using background knowledge in ontology matching*

2008-23 Stefan Visscher (UU), *Bayesian network models for the management of ventilator-associated pneumonia*

2008-22 Henk Koning (UU), *Communication of IT-Architecture*

2008-21 Krisztian Balog (UVA), *People Search in the Enterprise*

2008-20 Rex Arendsen (UVA), *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven.*

2008-19 Henning Rode (UT), *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*

2008-18 Guido de Croon (UM), *Adaptive Active Vision*

2008-17 Martin Op 't Land (TUD), *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*

2008-16 Henriëtte van Vugt (VU), *Embodied agents from a user's perspective*

2008-15 Martijn van Otterlo (UT), *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.*

2008-14 Arthur van Bunningen (UT), *Context-Aware Querying; Better Answers with Less Effort*

2008-13 Caterina Carraciolo (UVA), *Topic Driven Access to Scientific Handbooks*

2008-12 József Farkas (RUN), *A Semiotically Oriented Cognitive Model of Knowledge Representation*

2008-11 Vera Kartseva (VU), *Designing Controls for Network Organizations: A Value-Based Approach*

2008-10 Wauter Bosma (UT), *Discourse oriented summarization*

2008-09 Christof van Nimwegen (UU), *The paradox of the guided user: assistance can be counter-effective*

2008-08 Janneke Bolt (UU), *Bayesian Networks: Aspects of Approximate Inference*

2008-07 Peter van Rosmalen (OU), *Supporting the tutor in the design and support of adaptive e-learning*

2008-06 Arjen Hommersom (RUN), *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*

2008-05 Bela Mutschler (UT), *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*

2008-04 Ander de Keijzer (UT), *Management of Uncertain Data – towards unattended integration*
2008-03 Vera Hollink (UVA), *Optimizing hierarchical menus: a usage-based approach*
2008-02 Alexei Sharpanskykh (VU), *On Computer-Aided Methods for Modeling and Analysis of Organizations*
2008-01 Katalin Boer-Sorbán (EUR), *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*
2007-25 Joost Schalken (VU), *Empirical Investigations in Software Process Improvement*
2007-24 Georgina Ramírez Camps (CWI), *Structural Features in XML Retrieval*
2007-23 Peter Barna (TUE), *Specification of Application Logic in Web Information Systems*
2007-22 Zlatko Zlatev (UT), *Goal-oriented design of value and process models from patterns*
2007-21 Karianne Vermaas (UU), *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*
2007-20 Slinger Jansen (UU), *Customer Configuration Updating in a Software Supply Network*
2007-19 David Levy (UM), *Intimate relationships with artificial partners*
2007-18 Bart Orriëns (UvT), *On the development an management of adaptive business collaborations*
2007-17 Theodore Charitos (UU), *Reasoning with Dynamic Networks in Practice*
2007-16 Davide Grossi (UU), *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*
2007-15 Joyca Lacroix (UM), *NIM: a Situated Computational Memory Model*
2007-14 Niek Bergboer (UM), *Context-Based Image Analysis*
2007-13 Rutger Rienks (UT), *Meetings in Smart Environments; Implications of Progressing Technology*
2007-12 Marcel van Gerven (RUN), *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*
2007-11 Natalia Stash (TUE), *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*
2007-10 Huib Aldewereld (UU), *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*
2007-09 David Mobach (VU), *Agent-Based Mediated Service Negotiation*
2007-08 Mark Hoogendoorn (VU), *Modeling of Change in Multi-Agent Organizations*
2007-07 Nataša Jovanović (UT), *To Whom It May Concern – Addressee Identification in Face-to-Face Meetings*
2007-06 Gilad Mishne (UVA), *Applied Text Analytics for Blogs*
2007-05 Bart Schermer (UL), *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*
2007-04 Jurriaan van Diggelen (UU), *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*
2007-03 Peter Mika (VU), *Social Networks and the Semantic Web*
2007-02 Wouter Teepe (RUG), *Reconciling Information Exchange and Confidentiality: A Formal Approach*
2007-01 Kees Leune (UvT), *Access Control and Service-Oriented Architectures*
2006-28 Börkur Sigurbjörnsson (UVA), *Focused Information Access using XML Element Retrieval*
2006-27 Stefano Bocconi (CWI), *Vox Populi: generating video documentaries from semantically annotated media repositories*
2006-26 Vojkan Mihajlović (UT), *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*
2006-25 Madalina Drugan (UU), *Conditional log-likelihood MDL and Evolutionary MCMC*
2006-24 Laura Hollink (VU), *Semantic Annotation for Retrieval of Visual Resources*
2006-23 Ion Juvina (UU), *Development of Cognitive Model for Navigating on the Web*
2006-22 Paul de Vrieze (RUN), *Fundaments of Adaptive Personalisation*
2006-21 Bas van Gils (RUN), *Aptness on the Web*
2006-20 Marina Velikova (UvT), *Monotone models for prediction in data mining*
2006-19 Birna van Riemsdijk (UU), *Cognitive Agent Programming: A Semantic Approach*
2006-18 Valentin Zhizhkun (UVA), *Graph transformation for Natural Language Processing*
2006-17 Stacey Nagata (UU), *User Assistance for Multitasking with Interruptions on a Mobile Device*
2006-16 Carsten Riggelsen (UU), *Approximation Methods for Efficient Learning of Bayesian Networks*
2006-15 Rainer Malik (UU), *CONAN: Text Mining in the Biomedical Domain*
2006-14 Johan Hoorn (VU), *Software Requirements: Update, Upgrade, Redesign – towards a Theory of Requirements Change*
2006-13 Henk-Jan Lebbink (UU), *Dialogue and Decision Games for Information Exchanging Agents*
2006-12 Bert Bongers (VU), *Interactivation – Towards an e-cology of people, our technological environment, and the arts*
2006-11 Joeri van Ruth (UT), *Flattening Queries over Nested Data Types*
2006-10 Ronny Siebes (VU), *Semantic Routing in Peer-to-Peer Systems*
2006-09 Mohamed Wahdan (UM), *Automatic Formulation of the Auditor's Opinion*
2006-08 Eelco Herder (UT), *Forward, Back and Home Again – Analyzing User Behavior on the Web*
2006-07 Marko Smiljanic (UT), *XML schema matching – balancing efficiency and effectiveness by means of clustering*
2006-06 Ziv Baida (VU), *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling*
2006-05 Cees Pierik (UU), *Validation Techniques for Object-Oriented Proof Outlines*
2006-04 Marta Sabou (VU), *Building Web Service Ontologies*
2006-03 Noor Christoph (UVA), *The role of metacognitive skills in learning to solve problems*
2006-02 Cristina Chisalita (VU), *Contextual issues in the design and use of information technology in organizations*
2006-01 Samuil Angelov (TUE), *Foundations of B2B Electronic Contracting*

2005-21 Wijnand Derks (UT), *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*
2005-20 Cristina Coteanu (UL), *Cyber Consumer Law, State of the Art and Perspectives*
2005-19 Michel van Dartel (UM), *Situated Representation*
2005-18 Danielle Sent (UU), *Test-selection strategies for probabilistic networks*
2005-17 Boris Shishkov (TUD), *Software Specification Based on Re-usable Business Components*
2005-16 Joris Graaumans (UU), *Usability of XML Query Languages*
2005-15 Tibor Bosse (VU), *Analysis of the Dynamics of Cognitive Processes*
2005-14 Borys Omelayenko (VU), *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*
2005-13 Fred Hamburg (UL), *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*
2005-12 Csaba Boer (EUR), *Distributed Simulation in Industry*
2005-11 Elth Ogston (VU), *Agent Based Matchmaking and Clustering – A Decentralized Approach to Search*
2005-10 Anders Bouwer (UVA), *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*
2005-09 Jeen Broekstra (VU), *Storage, Querying and Inferencing for Semantic Web Languages*
2005-08 Richard Vdovjak (TUE), *A Model-driven Approach for Building Distributed Ontology-based Web Applications*
2005-07 Flavius Frasincar (TUE), *Hypermedia Presentation Generation for Semantic Web Information Systems*
2005-06 Pieter Spronck (UM), *Adaptive Game AI*
2005-05 Gabriel Infante-Lopez (UVA), *Two-Level Probabilistic Grammars for Natural Language Parsing*
2005-04 Nirvana Meratnia (UT), *Towards Database Support for Moving Object data*
2005-03 Franc Grootjen (RUN), *A Pragmatic Approach to the Conceptualisation of Language*
2005-02 Erik van der Werf (UM)), *AI techniques for the game of Go*
2005-01 Floor Verdenius (UVA), *Methodological Aspects of Designing Induction-Based Applications*
2004-20 Madelon Evers (Nyenrode), *Learning from Design: facilitating multidisciplinary design teams*
2004-19 Thijs Westerveld (UT), *Using generative probabilistic models for multimedia retrieval*
2004-18 Vania Bessa Machado (UvA), *Supporting the Construction of Qualitative Knowledge Models*
2004-17 Mark Winands (UM), *Informed Search in Complex Games*
2004-16 Federico Divina (VU), *Hybrid Genetic Relational Search for Inductive Learning*
2004-15 Arno Knobbe (UU), *Multi-Relational Data Mining*
2004-14 Paul Harrenstein (UU), *Logic in Conflict. Logical Explorations in Strategic Equilibrium*
2004-13 Wojciech Jamroga (UT), *Using Multiple Models of Reality: On Agents who Know how to Play*
2004-12 The Duy Bui (UT), *Creating emotions and facial expressions for embodied agents*
2004-11 Michel Klein (VU), *Change Management for Distributed Ontologies*
2004-10 Suzanne Kabel (UVA), *Knowledge-rich indexing of learning-objects*
2004-09 Martin Caminada (VU), *For the Sake of the Argument; explorations into argument-based reasoning*
2004-08 Joop Verbeek (UM), *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politile gegevensuitwisseling en digitale expertise*
2004-07 Elise Boltjes (UM), *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*
2004-06 Bart-Jan Hommes (TUD), *The Evaluation of Business Process Modeling Techniques*
2004-05 Viara Popova (EUR), *Knowledge discovery and monotonicity*
2004-04 Chris van Aart (UVA), *Organizational Principles for Multi-Agent Architectures*
2004-03 Perry Groot (VU), *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*
2004-02 Lai Xu (UvT), *Monitoring Multi-party Contracts for E-business*
2004-01 Virginia Dignum (UU), *A Model for Organizational Interaction: Based on Agents, Founded in Logic*
2003-18 Levente Kocsis (UM), *Learning Search Decisions*
2003-17 David Jansen (UT), *Extensions of Statecharts with Probability, Time, and Stochastic Timing*
2003-16 Menzo Windhouwer (CWI), *Feature Grammar Systems – Incremental Maintenance of Indexes to Digital Media Warehouses*
2003-15 Mathijs de Weerdt (TUD), *Plan Merging in Multi-Agent Systems*
2003-14 Stijn Hoppenbrouwers (KUN), *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*
2003-13 Jeroen Donkers (UM), *Nosce Hostem – Searching with Opponent Models*
2003-12 Roeland Ordelman (UT), *Dutch speech recognition in multimedia information retrieval*
2003-11 Simon Keizer (UT), *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*
2003-10 Andreas Lincke (UvT), *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*
2003-09 Rens Kortmann (UM), *The resolution of visually guided behaviour*
2003-08 Yongping Ran (UM), *Repair Based Scheduling*
2003-07 Machiel Jansen (UvA), *Formal Explorations of Knowledge Intensive Tasks*
2003-06 Boris van Schooten (UT), *Development and specification of virtual environments*
2003-05 Jos Lehmann (UVA), *Causation in Artificial Intelligence and Law – A modelling approach*
2003-04 Milan Petković (UT), *Content-Based Video Retrieval Supported by Database Technology*
2003-03 Martijn Schuemie (TUD), *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*
2003-02 Jan Broersen (VU), *Modal Action Logics for Reasoning About Reactive Systems*
2003-01 Heiner Stuckenschmidt (VU), *Ontology-Based Information Sharing in Weakly Structured Environments*
2002-17 Stefan Manegold (UVA), *Understanding, Modeling, and Improving Main-Memory Database Performance*

2002-16 Pieter van Langen (VU), *The Anatomy of Design: Foundations, Models and Applications*
2002-15 Rik Eshuis (UT), *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*
2002-14 Wieke de Vries (UU), *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*
2002-13 Hongjing Wu (TUE), *A Reference Architecture for Adaptive Hypermedia Applications*
2002-12 Albrecht Schmidt (Uva), *Processing XML in Database Systems*
2002-11 Wouter C.A. Wijngaards (VU), *Agent Based Modelling of Dynamics: Biological and Organisational Applications*
2002-10 Brian Sheppard (UM), *Towards Perfect Play of Scrabble*
2002-09 Willem-Jan van den Heuvel (KUB), *Integrating Modern Business Applications with Objectified Legacy Systems*
2002-08 Jaap Gordijn (VU), *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*
2002-07 Peter Boncz (CWI), *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*
2002-06 Laurens Mommers (UL), *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*
2002-05 Radu Serban (VU), *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*
2002-04 Juan Roberto Castelo Valdueza (UU), *The Discrete Acyclic Digraph Markov Model in Data Mining*
2002-03 Henk Ernst Blok (UT), *Database Optimization Aspects for Information Retrieval*
2002-02 Roelof van Zwol (UT), *Modelling and searching web-based document collections*
2002-01 Nico Lassing (VU), *Architecture-Level Modifiability Analysis*
2001-11 Tom M. van Engers (VUA), *Knowledge Management: The Role of Mental Models in Business Systems Design*
2001-10 Maarten Sierhuis (UvA), *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design*
2001-09 Pieter Jan 't Hoen (RUL), *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*
2001-08 Pascal van Eck (VU), *A Compositional Semantic Structure for Multi-Agent Systems Dynamics.*
2001-07 Bastiaan Schonhage (VU), *Diva: Architectural Perspectives on Information Visualization*
2001-06 Martijn van Welie (VU), *Task-based User Interface Design*
2001-05 Jacco van Ossenbruggen (VU), *Processing Structured Hypermedia: A Matter of Style*
2001-04 Evgueni Smirnov (UM), *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*
2001-03 Maarten van Someren (UvA), *Learning as problem solving*
2001-02 Koen Hindriks (UU), *Agent Programming Languages: Programming with Mental Models*
2001-01 Silja Renooij (UU), *Qualitative Approaches to Quantifying Probabilistic Networks*
2000-11 Jonas Karlsson (CWI), *Scalable Distributed Data Structures for Database Management*
2000-10 Niels Nes (CWI), *Image Database Management System Design Considerations, Algorithms and Architecture*
2000-09 Florian Waas (CWI), *Principles of Probabilistic Query Optimization*
2000-08 Veerle Coupé (EUR), *Sensitivity Analyis of Decision-Theoretic Networks*
2000-07 Niels Peek (UU), *Decision-theoretic Planning of Clinical Patient Management*
2000-06 Rogier van Eijk (UU), *Programming Languages for Agent Communication*
2000-05 Ruud van der Pol (UM), *Knowledge-based Query Formulation in Information Retrieval.*
2000-04 Geert de Haan (VU), *ETAG, A Formal Model of Competence Knowledge for User Interface Design*
2000-03 Carolien M.T. Metselaar (UVA), *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.*
2000-02 Koen Holtman (TUE), *Prototyping of CMS Storage Management*
2000-01 Frank Niessink (VU), *Perspectives on Improving Software Maintenance*
1999-08 Jacques H.J. Lenting (UM), *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.*
1999-07 David Spelt (UT), *Verification support for object database design*
1999-06 Niek J.E. Wijngaards (VU), *Re-design of compositional systems*
1999-05 Aldo de Moor (KUB), *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*
1999-04 Jacques Penders (UM), *The practical Art of Moving Physical Objects*
1999-03 Don Beal (UM), *The Nature of Minimax Search*
1999-02 Rob Potharst (EUR), *Classification using decision trees and neural nets*
1999-01 Mark Sloof (VU), *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*
1998-05 E.W. Oskamp (RUL), *Computerondersteuning bij Straftoemeting*
1998-04 Dennis Breuker (UM), *Memory versus Search in Games*
1998-03 Ans Steuten (TUD), *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*
1998-02 Floris Wiesman (UM), *Information Retrieval by Graphically Browsing Meta-Information*
1998-01 Johan van den Akker (CWI), *DEGAS – An Active, Temporal Database of Autonomous Objects*