Project no.: **IST-FP6-IP-027657**

Project full title: **Perception, Action & Cognition through Learning of Object-Action Complexes**

Project Acronym: **PACO-PLUS**

Deliverable no.: D3.1.1

Title of the deliverable: Meaning of Action

Contractual Date of Delivery to the CEC:	**31st January 2007**
Actual Date of Delivery to the CEC:	**31st January 2007**
Organisation name of lead contractor for this deliverable:	**Aalborg University (AAU)**
Author(s): Volker Krüger, Danica Kragic, Ales Ude and Christopher Geib	
Participants(s): AAU, KTH, UniKarl, BCCN, JSI, UEDIN	
Work package contributing to the deliverable:	**WP3.1**
Nature:	**R**
Version:	**1.0**
Total number of pages:	**28**
Start date of project:	1st Feb. 2006 **Duration:** 48 month

Projectco-funded by the European Commission within the Sixth Framework Programme (2002-2006) Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract:

This report contains the Technical Report as promised under deliverable D3.1.1. This report has been submitted to the Int. Journal on Advanced Robotics.

In this paper, we analyze the different approaches taken to-date within the computer vision, the robotics and the artificial intelligence community for the representation, recognition, synthesis and understanding of action. We present a common framework for dealing with *action* at different levels of complexity and provide the reader with the necessary related literature references. We put the literature reference further into context and outline a unified interpretation of action by taking into account the different aspects from robotics, vision and AI.

Keyword list: action recognition, action representation, computer vision, robotics, AI

AALBORG UNIVERSITY

Copenhagen Institute of Technology

Computer Vision and Machine Intelligence Lab (CVMI)

The Meaning of Action
A review on action recognition and mapping

Volker Krüger, Danica Kragic, Aleš Ude, Christopher Geib

Advances in Computer Vision and Machine Intelligence CVMI 2007:1

ISSN 1902-2034

Volker Krüger, Danica Kragic, Aleš Ude, Christopher Geib
The Meaning of Action
A review on action recognition and mapping

Report number: CVMI 2007:1
ISSN 1902-2034
Publication date: January 2007
E-mail of author: vok@media.aau.dk, dani@kth.se, ales.ude@ijs.si, c.geib@ed.ac.uk

Reports can be ordered from:

Copenhagen Institute of Technology
Aalborg University (AAU)
DK-2970 Ballerup
DENMARK

telefax: +45 96 35 24 80
http://www.cvmi.aau.dk/

The Meaning of Action
A review on action recognition and mapping

Volker Krüger	Danica Kragic	Aleš Ude	Christopher Geib
Aalborg Media Lab	Computer Vision and	Jozef Stefan Institute	School of Informatics
Aalborg University	Active Perception Lab	Dept. of Automatics,	Univ. of Edinburgh
Ballerup, Denmark	CSC-KTH	Biocyb. & Robotics	Edinburgh, Scotland
	Stockholm, Sweden	Ljubljana, Slovenia	

January 8, 2007

Abstract

In this paper, we analyze the different approaches taken to-date within the computer vision, the robotics and the artificial intelligence community for the representation, recognition, synthesis and understanding of action. We present a common framework for dealing with *action* at different levels of complexity and provide the reader with the necessary related literature references. We put the literature reference further into context and outline a unified interpretation of action by taking into account the different aspects from robotics, vision and AI.

Keywords: action recognition, action representation, computer vision, robotics, AI

1 Introduction

The recognition and interpretation of human or robot induced actions and activities has gained considerable interest in the computer vision, robotics and AI communities. This is partially due to increasing computer power that allows large amount of input data to be stored and processed, but also due the large number of potential applications, e.g., in visual surveillance, in the entertainment industry and for robot control. Depending on the application, starting points and aims in action based research are different. In this paper, we analyze the different approaches to action recognition and mapping taken to-date within the three communities.

In visual surveillance, many applications are limited to distinguish usual from unusual actions, without any further interpretation of the action in the scene. An application of great potentials is an automatic scene understanding system that includes the interpretation of the observed actions such as *what* actions are executed, *where* they are executed, *who* is involved, and even a *prediction* of what the observed individuals' intentions might be given their present behavior. Such a surveillance system has to be non-intrusive and could potentially include a number of different sensors. In the entertainment industry, the interest lies mainly in the field of motion capture and synthesis. In film productions, precise motion capture allows to replace an actor with a digital avatar (as often done in recent movies). In computer games, game designers are interested in realistically looking digital animations as well as in motion capture technology that allows the gamer to interact with the computer game through body movements, as, e.g., done in the Sony EyeToy games. Ideally, the motion capture should be non-intrusive for both, film and computer games, so that actors and gamers would not need to wear special suits. The computer game needs to be able to interpret the movements of the gamer in a

robust and reliable manner to maintain a maximal degree of entertainment. The surveillance and entertainment applications receive a strong attention from the computer vision community. Here, action recognition is often treated as a pattern matching problem with an additional time-dimension. A strong attention is focused on improper imaging conditions, noisy input data and the development of robust approaches for representing and recognizing the actions.

There is strong neurobiological evidence that human actions and activities are directly connected to the motor control of the human body [43, 94, 95]. When viewing other agents performing an action, the human visual system seems to relate the visual input to a sequence of motor primitives. The neurobiological representation for visually perceived, learned and recognized actions appears to be the same as the one used to drive the motor control of the body. These findings have gained considerable attention from the robotics community [23, 101] where the goal of *imitation learning* is to develop robot systems that are able to relate perceived actions of another (human) agent to its own embodiment in order to learn and later to recognize and to perform the demonstrated actions. Here, action representations based on detailed human body models are usually applied.

In robotics as well as in vision, the neurobiological findings motivate research to identify a set of action primitives that allow a) representation of the visually perceived action and b) motor control for imitation. In addition, this gives rise to the idea of interpreting and recognizing activities in a video scene through a hierarchy of primitives, simple actions and activities. Many researchers in vision and robotics attempt to learn the action or motor primitives by defining a "suitable" representation and then learning the primitives from demonstrations. The representations used to describe the primitives vary a lot across the literature and are subject to ongoing research.

As an example, for imitation learning a teacher might attempt to show a robot how to set-up or clean a dinner table. An important aspect is that the setting of the environment might change between the demonstration and the execution time. A robot that has to set-up a dinner table may have to plan the order of handling plates, cutlery and glasses in a different way than previously demonstrated by the human teacher. Hence, it is usually not sufficient to just replicate the human movements. Instead, the robot must have the ability to recognize what parts of the whole task can be segmented and considered as subtasks so that it can perform on-line planning for task execution given the current state of the environment. A number of crucial problems arise:

1. How should the robot be instructed that the temporal order of the subtasks may or may not matter? As an example, the main dish plate should always be under the appetizer plate while the temporal order in which the silverware is placed on the table is not important.

2. How should the scene, the objects and the changes that can be done to them be represented? For example, when cleaning up the table the representation should allow to pile on the tray wine glasses on top of plates while piling plates on wine glasses might cause a major disaster.

3. Given a specific scene state, it is not clear if the robot is actually able to perform a particular action. For example, the representation may specify that wine glasses *can* be piled on top of plates while, at the same time, the robot might not be tall enough to do so.

4. The entire scene may change during the planning phase so that the robot needs to be able to react to sudden scene changes.

The above problems have been considered in the area of AI in regard to task planning and sequencing with the specific focus on structured collections of actions. In other words, this area has

concentrated on development of different types of reasoning systems such as rule based systems, traditional Bayes nets, context free grammars, etc. mainly for task planning purposes. Different methods of action representation make the strongest obstacle to integrating the requirements in the robotics area with the solutions provided by the AI. In robotics, representations have to model continuous data while AI builds upon discrete symbolic representations. Hence, while AI representations focus mainly on modeling the high-level conceptual state changes that result from action execution, robotics community considers primarily the low-level continuous action execution.

In spite of the differences in the potential applications in the areas, most of the scenarios are closely related: all of them use sensory input, all need to capture the movements of an agent (at different degrees of precision) and all require a certain level of intelligence to *understand* the meaning of the captured movements. Thus there is a need to:

1. recognize the movements and actions of observed agents (*recognizing the action by observing it*)
2. understand what effects certain actions have on the environment of the actor (*recognizing the action by observing its effects on the environment*)
3. understand *how* to physically perform a certain action in order to *cause* a particular change in the environment.

While the first two points are commonly shared across members of a society (non-verbal communication), the third point depends heavily on the individual/robot under consideration: how to perform an action that causes a particular environmental change may be different between individuals and robots, e.g., depending on their physical conditions. Thus, the understanding of actions is in the gray zone between vision, robotics and artificial intelligence.

In this paper, we analyze the different approaches taken to-date within these three communities. We present a common framework for dealing with *action* at different levels of complexity and provide the reader with the necessary related literature references. Different authors use different terms for discussing action primitives and action grammars. In Sec. 2, we mention the most general references and define, to escape the diversity of terms, our own terminology that we will use throughout this paper. In Sec. 3, 4, 5 we discuss how the representation and recognition of actions is treated in the different communities. We conclude this paper in Sec. 6 with outlining a unified interpretation of action by taking into account aspects from robotics, vision and AI.

2 Notation and Action Hierarchies

Terms like *actions*, *activities*, *complex actions*, *simple actions* and *behaviors* are often used interchangeably by different authors. However, in order to describe and compare the different publications, we shortly review the different terms used and define a common terminology used throughout the paper. In a pioneering work [76], Nagel suggested to use a hierarchy of *change, event, verb, episode, history*. An alternative hierarchy (reflecting the computational aspects) is proposed by Bobick [12] who suggests to use *movement, activity and action* as different levels of abstraction (see also [1]). Others suggest to also include *situations* [45] or use a hierarchy of *Action primitives* and *Parent Behaviors* [53].

In this paper, we adopt the following action hierarchy: *action/motor primitives*, *actions* and *activities*. *Action primitives* or *motor primitives* are used for atomic entities out of which actions are built. *Actions* are, in turn, composed into *activities*. The granularity of the primitives often depends on the application. For example, in robotics, *motor primitives* are often understood as sets of motor control commands that are used to generate an action by the robot (see Sec. 3.4).

As an example, in tennis *action primitives* could be, e.g., "forehand", "backhand", "run left", "run right". The term *action* is used for a sequence of action primitives needed to return a ball. The choice of a particular action depends on whether a forehand, backhand, lob or volley etc., is required in order to be able to return the ball successfully. Most of the research discussed below falls into this category. The *activity* then is in this example "playing tennis". *Activities* are larger scale events that typically depend on the context of the environment, objects or interacting humans.

A good overview of activity recognition is given by Aggarwal and Park [1] and in a more recent one by Moeslund *et al.*[74]. They aim at higher-level understanding of activities and interactions and discuss different aspect such as level of detail, different human models, recognition approaches and high-level recognition schemes. Veeraraghavan *et al.*[121] discuss the structure of an action and activity space.

3 Interpretation and Recognition of Action in Computer Vision

The vision community has mainly the goal of detecting, recognizing and interpreting movements of a (possibly non-human) agent based on video camera data. For example, in scene interpretation for surveillance the knowledge is often represented in a statistical manner. It is meant to distinguish "regular" from "irregular" activities and it should be independent from the objects causing the activity and thus are usually not meant to distinguish explicitly, e.g, cars from humans. On the other hand, some surveillance applications focus explicitly on human activities and the interactions between them. Here, one finds both, holistic approaches, that take into account the entire human body without considering particular body parts, and human body model-based approaches that attempt to align a detailed human model to the observed video data. Most holistic approaches attempt to identify "holistic" information such as gender, identity or simple actions like walking or running. Researchers using human body model-based approaches appear often to be interested in more subtle actions or attempt to model actions by looking for action primitives with which the complex actions can be modeled. Local approaches can also be used in medical applications or in applications from the entertainment industry. In the following, we review some of the recent approaches.

3.1 Scene Interpretation

Many approaches consider the camera view as a whole and attempt to learn and recognize activities simply by observing the motion of objects without necessarily knowing their identity. This is reasonable in situations where the objects are small enough to be represented as points on a 2D plane.

Stauffer *et al.*[110] present a full scene interpretation system which allows detection of unusual situations. The system extracts features such as 2-D position and speed, size and binary silhouettes. Vector Quantization is applied to generate a codebook of K prototypes. Instead of taking the explicit temporal relationship between the symbols into account, Stauffer and Grimson use co-occurrence statistics. Then, they define a binary tree structure by recursively defining two probability mass functions across the prototypes of the code book that best explain the co-occurrence matrix. The leaf nodes of the binary tree are probability distributions of co-occurrences across the prototypes and at a higher tree depth define simple scene activities like pedestrian and car movement. These can then be used for scene interpretation. Boiman and Irani [14] approach the problem of detection irregularities in a scene as a problem of composing newly observed data using spatio-temporal patches extracted from previously seen visual examples. They extract small image and video patches which are used as local descriptors. In an inference process, they search for patches with a similar geometric configuration and appearance

properties, while allowing for small local misalignments in their relative geometric arrangement. This way, they are able to quickly and efficiently infer subtle but important local changes in behavior.

In [21, 119] activity trajectories are modeled using non-rigid shapes and a dynamic model that characterizes the variations in the shape structure. Vaswani *et al.*[119] uses Kendall's statistical shape theory [60]. Nonlinear dynamical models are used to characterize the shape variation over time. An activity is recognized if it agrees with the learned parameters of the shape and associated dynamics. Chowdhury *et al.*[21] use a subspace method to model activities as a linear combination of 3D basis shapes. The work is based on the factorization theorem [116]. Deviations from the learned normal activity shapes can be used to identify abnormal ones. A similar complex task is approached by Xiang and Gong [128]. They present a unified bottom-up and top-down approach to model complex activities of multiple objects in cluttered scenes. Their approach is object-independent and they use a Dynamically Multi-Linked Hidden Markov Models (HMMs) in conjunction with Schwarz's Bayesian Information Criterion [104] to interlink between multiple temporal processes corresponding to multiple event classes.

3.2 Holistic Recognition Approaches

A large number of approaches for recognition are based on the human silhouette as whole silhouettes can often be extracted much easier when singular body parts are difficult to distinguish. This is especially true when the observed agent is far away from the camera. Naturally, the question on what an observed agent is precisely doing can be answered only with a much lesser precision than when singular body parts are extracted. Actions such as walking, running, jumping, etc. as well as their speed, location in the image and their direction can, however, be extracted with an impressive robustness.

All the approaches mentioned in this section attempt to recognize the apparent action based directly on a sequence of 2D image projections, without the intermediate use, e.g., of 3D human model. The argument is that the use of an explicit human (not necessarily 3D) model is often not feasible in case of noisy and imperfect imaging conditions and that a direct pattern recognition based on the 2D data is potentially more robust. This argument holds especially when there are only very few pixels on image of the observed agent.

A pioneering work has been presented by Efros *et al.*[29]. They attempt to recognize a set of simple actions (walking, running plus direction and location) of people whose images in the video are only 30 pixels tall and where the video quality is poor. They use a set of features that are based on blurred optic flow (blurred motion channels). First, the person is tracked so that the image is stabilized in the middle of a tracking window. The blurred motion channels are computed on the residual motion that is due to the motion of the body parts. Spatio-temporal cross-correlation is used for matching with a database. Roh *et al.*[97] base their action recognition task on curvature scale space templates of an agent's silhouette.

The work of Robertson and Reid [96] extends the work of Efros [29] by proposing an approach where complex actions can be dynamically composed out of the set of simple actions. They attempt to *understand* actions by building a hierarchical system that is based on reasoning with belief networks and HMMs on the highest level and on the lowest level with features such as position and velocity as action descriptors. The system is able to output qualitative information such as *walking – left-to-right – on the sidewalk*. To our knowledge, this is one of the very few papers that attempts to connect computer vision techniques with AI techniques in the context of action recognition.

A large number of publications work with space-time volumes which is a recently proposed representation for the spatio-temporal domain. The 3D contour of a person gives rise to a 2D projection. Considering this projection over time defines the XYT image volume. One

of the main ideas here is to use spatio-temporal XT-slices from an image volume XYT [91, 93]. Articulated motions of a human then show a typical trajectory pattern. Ricquebourg and Bouthemy [91] demonstrate how XT-slices can facilitate tracking and reconstruction of 2D motion trajectories. The reconstructed trajectory allows a simple classification between pedestrians and vehicles. Ritscher *et al.*[93] discuss the recognition in more detail by a closer investigation of the XT-slices. Quantifying the braided pattern in the slices of the spatio-temporal cube gives rise to a set of features (one for each slice) and their distribution is used to classify the actions.

Another approach is that of "Actions Sketches" or "Space-Time Shapes" in the 3D XYT volume. Yilmaz and Shah [131] extract information such as speed, direction and shape by analyzing the differential geometric properties of the XYT volume. They approach action recognition as an object matching task by interpreting the XYT as rigid 3D objects. Blank *et al.*[11] also analyze the XYT volume. They generalize techniques for the analysis of 2D shapes [46] for the use on the XYT volume. Blank *et al.*argue that the time domain introduces properties that do not exist in the xy-domain and needs thus a different treatment. For their analysis they utilize properties of the solution of the Poisson equation [46]. This gives rise to local and global descriptors that are used for recognizing simple actions.

Instead of using spatio-temporal volumes, a large number of researchers choose the more classical approach of considering sequences of silhouettes. Yu *et al.*[133] extract silhouettes and their contours are unwrapped and processed by PCA. A three-layer feed forward network is used to distinguish "walking", "running"and "other" based on the trajectories in eigenspace. In another PCA-based approach, Rahman and Robles-Kelly [86] suggest to use a tuned eigenspace technique. They tuned eigenspaces allow to treat the action problem as a nearest-neighborhood problem in eigenspace. Jiang *et al.*[55] attempt to match a given sequence of poses to a novel video. They treat this problem as an optimal matching problem by changing the usually highly non-convex problem in to a convex one.

Bobick and Davis pioneered the idea of temporal templates [12, 13]. They propose a representation and recognition theory [12, 13] that is based on *motion energy images* (MEI) and *motion history images* (MHI). The MEI is a binary cumulative motion image. The MHI is an enhancement of the MEI where the pixel intensities are a function of the motion history at that pixel. Matching temporal templates is based on Hu moments. Bradski *et al.*[15] pick up the idea of MHI and develop timed MHI (tMHI) for motion segmentation. tMHI allow determination of the normal optical flow. Motion is segmented relative to object boundaries and the motion orientation. Hu moments are applied to the binary silhouette to recognize the pose. Elgammal and Lee [2] use optic flow in addition to the shape features and a HMM is used to model the dynamics. In [32, 33], Elgammal and Lee use local linear embedding (LLE) [98, 115] in order to find a linear embedding of human silhouettes. In conjunction with a generalized radial basis function interpolation, they are able to separate style and content of the performed actions [33] as well as to infer 3D body pose from 2D silhouettes [32]. Sato and Aggarwal [100] are concerned with the detection of interaction between two individuals. This is done by grouping foreground pixels according to similar velocities. A subsequent tracker tracks the velocity blobs. The distance between two people, the slope of relative distance and the slope of each person's position are the features used for interaction detection and classification. Gao *et al.*[40] consider a smart room application. A dining room activity analysis is performed by combining motion segmentation with tracking. They use motion segmentation based on optical flow and RANSAC. Then, they combine the motion segmentation with a tracking approach which is sensitive to subtle motion. In order to identify activities, they identify predominant directions of relative movements.

In a number of publications, recognition is based on HMMs and dynamic Bayes networks (DBNs). The work of Yamato *et al.*[130] is an example of an early application of HMMs to the problem of action recognition. They demonstrated the usefulness of HMMs for the recognition

of sport scenes. Elgammal *et al.*[34] propose a variant of semi-continuous HMMs for learning gesture dynamics. They represent the observation function of the HMM as non-parametric distributions to be able to relate a large number of exemplars to a small set of states. Luo *et al.*[68] present a scheme for video analysis and interpretation where the higher-level knowledge and the spatio-temporal semantics of objects are encoded with DBNs. The DBNs are based on key-frames and are defined for video objects. Shi *et al.*[106] present an approach for semi-supervised learning of the HMM or DBN states to incorporate prior knowledge.

3.3 Recognition Based on Body Parts

Despite the concerns mentioned in Sec. 3.2 about the difficulties in detecting singular body parts, many authors are concerned with the recognition of actions based on the dynamics and settings of individual body parts. Some approaches, e.g., [26], start out with silhouettes and detect the body parts using a method inspired by the W4-system [48] which seems to work well under the assumption of good foreground-background separation and large enough number of pixels on the observed agent. Other authors use 3D-model based body tracking approaches where the recognition of (periodic) action is used as a loop-back to support pose estimation [81, 107, 27, 7]. Many authors attempt to consider the problem of detecting body parts and recognizing actions as a joint problem by defining the action representation strictly based on the data that *can* be extracted [47, 105, 37, 132]. Other approaches circumvent the vision problem by using a motion capture system in order to be able to focus on finding good representations of actions [25, 83].

In a work related to [125], Wang *et al.*[124] present an approach where contours are extracted and a mean contour is computed to represent the static contour information. Dynamic information is extracted by using a detailed model composed of 14 rigid body parts, each one represented by a truncated cone. Particle filtering is used to compute the likelihood of a pose given an input image.

Ren and Xu [89] use as input a binary silhouette from which they detect the head, torso, hands and elbow angles. Then, a primitive-based coupled HMM is used to recognize natural complex and predefined actions. They extend their work in [90] by introducing primitive-based DBNs. Parameswaran and Chellappa [83] consider the problem of view-invariant action recognition based on point-light displays by investigating 2D and 3D invariant theory. As no general, non-trivial 3D-2D invariants exist, Parameswaran and Chellappa employ a convenient 2D invariant representation by decomposing and combining the patches of a 3D scene. For example, key poses can be identifies where joints in the different poses are aligned. In the 3D case, six-tuples corresponding to six joints give rise to 3D invariant values and it is suggested to use the progression of these invariants over time for action representation. A similar issue is discussed in the work by Yilmaz and Shah [132] where joint trajectories from several uncalibrated moving cameras are considered. They propose an extension to the standard epipolar geometry based approach by introducing a temporal fundamental matrix that models the effects of the camera motion. The recognition problem is then approached in terms of the quality of the recovered scene geometry. Gritai *et al.*[47] address the invariant recognition of human actions, and investigate the use of anthropometry to provide constraints on matching. Gritai *et al.*use the constraints to measure the similarity between poses and pose sequences. Their work is based on a point-light display like representation where a pose is presented through a set of points in 3D space. Sheikh *et al.*[105] pick up these results of [47, 132] and discuss that the three most important sources of variability in the task of recognizing actions come from variations in viewpoint, execution rate and anthropometry of the actors. Then, they argue that the variability associated with the execution of an action can be closely approximated by a linear combination of action bases in joint spatio-temporal space. Davis' and Gao's [24] aim is to recognize properties from visual target cues, e.g. the sex of an individual or the weight of a carried object is

estimated from how the individuals move. Labeled 2D trajectories from motion capture devices of humans are factored using three-mode PCA into components interpreted as *posture*, *time* and *effort*. An importance weight for each of the trajectories is learned automatically. In order to detect particular body parts Fanti *et al.*[37] give the structure of a human as model knowledge. To find the most likely model alignment with input data they exploit appearance information which remains approximately invariant within the same setting. Expectation maximization is used for unsupervised learning of the parameters and structure of the model for a particular action and unlabeled input data. Action is then recognized by maximum likelihood estimation.

3.4 Action Primitives and Grammars

Some of the work attempt to decouple actions into action primitives and to interpret actions as a composition on the alphabet of these action primitives, however, without the constraints of having to drive a motor controller with the same representation. E.g. Vecchio and Perona [120] employ techniques from the dynamical systems framework to approach segmentation and classification. System identification techniques are used to derive analytical error analysis and performance estimates. Once, the primitives are detected an iterative approach is used to find the sequence of primitives for a novel action. Lu *et al.*[67] also approach the problem from a system theoretic point of view. Their goal is to segment and represent repetitive movements. For this, they model the joint data over time with a second order auto-regressive (AR) model and the segmentation problem is approached by detection significant changes of the dynamical parameters. Then, for each motion segment and for each joint, they model the motion with a damped harmonic model. In order to compare actions, a metric based on the dynamic model parameters is defined. A different problem is studied by Wang *et al.*[123] addressing what kind of cost function should be used to assure smooth transitions between primitives.

While most scientists concentrate on the action representation by circumventing the vision problem, Rao *et al.*[88] take a vision-based approach. They propose a view-invariant representation of action based on *dynamic instants* and *intervals*. Dynamic instants (key poses) are used as primitives of actions which are computed from discontinuities of 2D hand trajectories. An interval represents the time period between two dynamic instants.

Modeling of activities on a semantic level has been attempted by Park and Aggarwal [84]. The system they describe has 3 abstraction levels. At the first level, human body parts are detected using a Bayesian network. At the second level, DBNs are used to model the actions of a single person. At the highest level, the results from the second level are used to identify the interactions between individuals. Ivanov and Bobick [52] suggest using stochastic parsing for a semantic representation of an action. They discuss that for some activities, where it comes to semantic or temporal ambiguities or insufficient data, stochastic approaches may be insufficient to model complex actions and activities. They suggest decoupling actions into primitive components and using a stochastic parser for recognition. In [52] they pick up a work by Stolcke [113] on syntactic parsing in speech recognition and enhance this work for activity recognition in video data. To be able to work with grammars, one needs to be able to decouple complex actions in to action primitives. Krger [62] suggests to embed the HMMs of different action primitives into a Bayesian framework over time which identifies, at each time instance, the most likely action primitive. Yamamoto *et al.*[129] present an application where a stochastic context free grammar is used for action recognition. A very interesting approach is presented by Lv and Nevatia in [69] where the authors are interested in recognizing and segmenting full-body human action. Lv and Nevatia decompose the large joint space into a set feature spaces where each feature corresponds to a single joint or combinations of related joints. They use then HMMs to recognize each action class based on the features and an AdaBoost scheme to detect and recognize the features.

4 Interpretation and Recognition of Action in Robotics

Unlike vision, robotics is mainly concerned with generative models of action. The robotics community has, however, recognized that the acquisition of new behaviors can be realized by observing and generalizing the behaviors of other agents. The combination of generative models and action recognition leads to robots that can imitate the behavior of other individuals [101, 16, 28].

Hence, the interest of roboticist is to enable robots with action recognition capabilities, both if these actions are performed by humans or other robots. In some cases, the action recognition is used for pure recognition purposes in context understanding or interaction. Consequently, different discriminative approaches are commonly adopted here. However, recent developments in the field of humanoid robots have motivated the use and investigation of generative approaches with the particular application of making robots move and excite their action in a *human-like* way, thus raising interest in integrated action recognition and action generation approaches.

For a robot that has to perform tasks in a human environment, it is also necessary to be able to learn about objects and object categories. It has been recognized recently that grounding in the embodiment of a robot, as-well as continuous learning is required to facilitate learning of objects and object categories [114, 38]. The idea is that robots will not be able to form useful categories or object representations by only being a passive observer of its environment. Rather a robot should, like a human infant, learn about objects by interacting with them, forming representations of the objects and their categories that are grounded in its embodiment. Most of the work on robotic grasping has been dealing with analytical methods where the shape of the objects being grasped is known *a-priori*. This problem is important and difficult mainly because of the high number of DOFs involved in grasping arbitrary objects with complex hands.

One of the most basic interactions that can occur between a robot and an object is for the robot to push the object, i.e. to simply make a physical contact. Already at this stage, the robot should be able to form two categories: physical and non-physical objects, where a physical object is categorized by the fact that interaction forces occur. A higher level interaction between the robot and an object would exist if the robot was able to *grasp* the object. In this case, the robot would gain actual physical control over the object and having the possibility to perform controlled actions on it, such as examining it from other angles, weighing it, placing it etc. Information obtained during this interaction can then be used to update the robots representations about objects and the world. Furthermore, the successfully performed grasps can be used as ground truth for future grasp refinement [38].

Some of the questions that are interesting for robotics:

- What modeling strategies are suitable for action representation and recognition purposes?
- Is it possible to learn action when we do not have the knowledge of the task or the embodiment (kinematic structure) of the teacher?
- Is it possible to distinguish between very similar actions such as *pick up* and *push* an object?
- Is it enough to only observe the motion of the arm/hand or does the motion of the object have to be included in the modeling process?

4.1 Movement Primitives in Robotics

Many of the generative approaches have found their roots in the work of Newton *et al.*[35] where the behavioral experiments indicated that observers are able to segment ongoing activity into temporal parts named *action units*. In addition, it has been shown that the resulting

segmentation is reliable and systematically related to relevant features of the action. Arbib [3] proposed the idea of movement primitives, which can be viewed as a sequence of actions that accomplish a complete goal-directed behavior. Conceptually, the idea of of movement primitives is appealing because it allows us to abstract complex motions as symbols, thus providing the basis for higher level cognitive processes. This has been demonstrated in [70], where motor behaviors execute the appropriate primitives to accomplish a verbally described high-level task.

There is no consensus in the literature about how to encode movement primitives (see also Sec. 3.4). Proposals include nonlinear dynamic attractor systems that can be flexibly adjusted to represent arbitrarily complex motor behaviors [102], primitive flow fields acquired from the motion capture data [54], hierarchical recurrent neural networks [82], HMMs [10, 51], and movement representation by force fields [75] . There may well be that no single representation exists and that different movement primitives are encoded differently.

More specifically, Jenkins *et al.*[54] suggest to apply a spatio-temporal non-linear dimension reduction technique on manually or automatically segmented human motion capture data. Similar segments are clustered into primitive units which are generalized into parameterized primitives by interpolating between them. In the same manner, they define action units ("behavior units") which can be generalized into actions. Ijspeert *et al.*[49, 102] define a set of nonlinear differential equations that form a control policy (CP) and quantify how well different trajectories can be fitted with these CPs. The parameters of a CP for a primitive movement are learned in a training phase. These parameters are also used to compute similarities between movements. Billard *et al.*[10] use an HMM based approach to learn characteristic features of repetitively demonstrated movements. They suggest to use the HMM to synthesize joint trajectories of a robot. For each joint, one HMM is used. Calinon *et al.*[18] use an additional HMM to model end-effector movement. In these approaches, the HMM structure is heavily constrained to assure convergence to a model that can be used for synthesizing joint trajectories. Paine and Tani [82] propose a hierarchical recurrent neural network that can both encode the sensorimotor primitives and switch between them. Different types of dynamic structures self-organize in the lower and higher levels of the network. The interplay of task-specific top-down and bottom-up processes allows the execution of complex navigation tasks.

This motivates the idea that – in view of imitation learning – the action recognition process may be considered as an interpretation of the continuous human behaviors which, in its turn, consists of a sequence of action primitives such as *reaching*, *picking up*, *putting down*. The key issues are how to identify what the movement primitives in a given domain are, how to encode them and how to recognize them in the motion capture data. Finally, imitation learning requires to relate movement primitives of other agents to the robot's own primitive movements. While many of the above mentioned approaches provide methods to learn the parameters of movement primitives in a given domain, the automatic determination of all relevant primitives in a domain has proven to be extremely difficult. They are therefore often hand designed [8] or acquired from the motion capture data with the help of manual segmentation.

4.2 Imitation Learning

We have seen that the integration of action recognition with generative models for movements and actions leads to imitation learning. It has been argued that imitation learning needs to address the following three questions: 1. what to imitate, 2. how to imitate, and 3. when to imitate [78]. The first issue is concerned with the perception of actions, the second with action generation and the third with decision making. In the following we review the works that are concerned with the first two issues.

Robotics research on imitation started in early 1990s under the names such as teaching by showing, learning by watching, and programming by demonstration. Roboticists first focused on

the extraction of the task knowledge by observing and analyzing the changes in the environment caused by a human performing an assembly task [50, 63]. Kuniyoshi *et al.*[63] and Kang and Ikeuchi [56] also stressed the importance of tracking and segmenting the demonstrator's hand motion to acquire additional information about the task. Thus, already from the beginning it became clear that imitation depends on the analysis and recognition of human motion, the identification of object configurations relevant to the task, and the detection of transitions between object configurations.

With the advent of humanoid robots, which have a kinematic and dynamic structure similar to humans, the acquisition of motor knowledge by observing humans performance has become more attractive. First works dealt with the mapping of human grasps to the grasps of a humanoid hand [57]. The mapping of whole body human movements, e. g. dance movements, to the movements of a humanoid robot followed [117, 92]. An automatic approach to relate human kinematics to humanoid robot kinematics has been developed [118] and it has been shown how to incorporate balancing controllers into the captured movements [99].

Kuniyoshi *et al.*[64] focus on the very basic question of how the robot can acquire the appearance-level imitation ability. They start from the proposal of Meltzoff and Moore [72] who found that very early neonates exhibit the imitation ability. Meltzoff and Moore proposed that either there exists an innate mechanism which represents the gestural mechanism or such a representation is built through self-exploratory sensory motor learning called body babbling. Kuniyoshi *et al.*[64] followed the second approach and created a humanoid that learns to imitate first-seen gestural movements by performing self-exploratory motion.

The appearance-level imitation of movements adapted to the robot kinematics and/or dynamics is often not sufficient to achieve the task goal. Many tasks require to consider the effect of movements on the target objects. Miyamoto *et al.*[73] extract a set of via-points from a human movement trajectory and treat the extracted via points as control variables to accomplish the task. Atkeson and Schaal [5] studied learning of motor tasks from human demonstration based on learning a task model and a reward function from the demonstration and use the model and reward function to compute an appropriate policy. Nakanishi *et al.*[77] introduced a framework for the learning of walking controllers using dynamic movement primitives. Asfour *et al.*[4] use HMMs to generalize movements demonstrated to a robot several times.

Yet a higher level of abstraction is achieved by sequencing a number of action units. HHMs have been proposed as a suitable representation for this purpose [79, 10, 51, 4]. These approaches attempt to integrate action recognition with movement generation. HMMs define a joint probability distribution over observations and state variables. For modeling of the observation process and enumerating all possible sequences of observations, it is commonly assumed that these are atomic and independent. This affects the inference problem which makes probabilistic models intractable for multiple overlapping features of the observation or complex dependencies of observations at multiple time steps. One of the solutions to this problem may be the use of discriminative models such as Conditional Random Fields [109].

Billard *et al.*[10] argue that the data used for imitation has statistical dependencies between the activities one wishes to model and that each activity has a rich set of features that can aid both the modeling and recognition process. They developed a general policy for learning the relevant features of an imitation task.

The discovery of mirror neurons, which fire both when the subject observes and when the subjects generates a specific behavior, has greatly influenced research in robot imitation. Inamura *et al.*[51] proposed a model in which movement primitives can be both recognized and generated using the same HMMs, thus realizing the mirror neuron idea on a humanoid robot.

4.3 Learning Actions from Multiple Demonstrations

An important issue to consider for robotic applications is that the initial task setting will change between the demonstration and execution time. A robot that has to set-up a dinner table may have to plan the order of handling plates, cutlery and glasses in a different way that previously demonstrated by a human teacher. Hence, it is not sufficient to just replicate the human movements but the robot i) must have the ability to recognize what parts of the whole task can be segmented and considered as subtasks so to ii) perform on-line planning for task execution given the current state of the environment. The important problem here is how to instruct or teach the robot the essential order of the subtasks for which the execution order may or may not be crucial. As an example, the main dish plate should always be under the appetizer or a soup plate and the order in which these are placed on the table is important. One way of addressing this problem is to demonstrate a task to the robot multiple times and let the robot learn which order of the subtasks is essential. Many of the current robot instruction systems concentrate on learning by imitation or PbD based on a single demonstration. However, the robot should be able to update the initial task model by observing humans or another robot performing the task. In other words, we need a task level learning system that builds constraints automatically identified from multiple demonstrations.

This problem has been studied by Ogawara *et al.*[61], where *essential interactions* are used to denote the important hand movements during an object manipulation task. Then, the relative trajectories corresponding to each essential interaction are generalized and stored in the task model, which is used to reproduce a skilled behavior. The work presented by Ekvall and Kragic [31] considers this problem not on the trajectory but on the task planning level where each demonstrated task is decomposed into subtasks that allow for segmentation and classification of the input data. The demonstrated tasks are then merged into a flexible task model, describing the task goal state and task constraints. The latter work is then also similar to the task level planning approaches studied in the field of artificial intelligence.

5 Representation and Recognition of Action in AI

In contrast to most of the work in this area from the robotics and vision side, the AI work in action and plan recognition has focused more on recognizing structured collections of actions. Traditionally this task has been called *plan recognition, task tracking*, or *intent recognition.* Sadly these terms in some cases have obscured the task that was actually being preformed. A great deal of research has been done on plan recognition using multiple approaches including: rule based systems, traditional Bayes nets, parsing of probabilistic (and non probabilistic) context free grammars, graph covering, and even marker passing. The rest of this discussion will be organized around the approaches used for plan recognition.

The earliest work in plan recognition [103, 126] was rule-based; researchers attempted to come up with inference rules that would capture the nature of plan recognition. However without an underlying formal model these rule sets are difficult to maintain and do not scale well. Later work [22] distinguish between two kinds of plan recognition *intended* and *keyhole*: In intended recognition, the agent is cooperative and its actions are done with the intend that they are understood. For example, a tutor demonstrating a procedure to a trainee would provide a case of intended recognition. In keyhole recognition, the recognizer is simply watching normal actions by an ambivalent agent. These cases arise, for example, in systems that are intended to watch some human user imperceptibly, and offer assistance, appropriate to context, when possible.

Kautz and Allen's early work [59] has framed much of the work in plan recognition to date. They defined the problem of keyhole plan recognition as a problem of identifying a minimal set of *top-level actions* sufficient to explain the set of observed actions. Plans were represented in a

plan graph, with top-level actions as root nodes and expansions of these actions into unordered sets of child actions representing plan decomposition. The problem of plan recognition was viewed as a problem of graph covering. Kautz and Allen formalized this in terms of McCarthy's circumscription [71].

Kautz also presented an approximate implementation of this approach that recasts the problem as one of computing vertex covers of the plan graph [58]. To gain efficiency, this implementation assumes that the observed agent is only attempting one top-level goal at a given time. Furthermore, it does not take into account differences in the *a priori* likelihood of different goals. Observing an agent going to the airport, this algorithm views "air travel," and "terrorist attack" as equally likely, since they both cover the observations.

Charniak and Goldman [19] argued that, plan recognition is just abduction, or reasoning to the best explanation [20], and it could therefore best be done as Bayesian (probabilistic) inference. This would support the preference for minimal explanations, in the case of equally likely hypotheses, but also correctly handle explanations of the same complexity but with different likelihoods. However, their system was unable to handle the case of failing to observe actions. Systems that observe the actual execution of actions, rather than consuming accounts thereof, often know that some actions have *not* been carried out and should be able to make use of this information. Neither Kautz and Allen nor Charniak and Goldman address this problem of evidence from failure to observe actions. For Charniak and Goldman, at least, this followed from their focus on plan recognition as part of story understanding. In human communication, stories are radically compressed by omitting steps that the reader or hearer can infer based on explicitly-mentioned material and background knowledge.

Systems like those of Charniak and Goldman and Kautz and Allen are not capable of reasoning like this, because they do not start from a model of plan execution over time. As a result, they cannot represent the fact that an action has not been observed *yet*. In general such systems take one of two solutions. First they can assert that the action has not and *will not* occur, or second they can be silent about whether an action has occurred — implying that the system has failed to notice the action, not that the action hasn't occurred. Both of these solutions are unsatisfying.

Both Vilain [122] and Sidner [108] present arguments for viewing plan recognition as parsing. The major problem with parsing as a model of plan recognition is that it does not treat partially-ordered plans or interleaved plans well. Both partial ordering and interleaving of plans require an exponential increase in the size of traditional context free grammars which can have a significant impact on the computational cost of the algorithm. There are grammatical formalisms that are powerful enough to capture interleaving. However, the advantage of parsing as a model is that it admits of efficient implementation when restricted to context-free languages. If this restriction is raised, this diminishes the argument for using parsing as a model.

Pynadath and Wellman [85] have proposed probabilistic parsing for plan recognition. Using plans represented as probabilistic context-free grammars (PCFGs) they build Bayes nets to evaluate observations. However, this approach still suffers from the problems of partial ordered and interleaved plans. They also propose that probabilistic context-*sensitive* grammars (PCSGs) might overcome this problem, but it is significantly more difficult to define a probability distribution for a PCSG.

Geib and Goldman [44, 42, 41] have presented a hybrid logical probabilistic plan recognition method that is based on weighted model counting. A complete and covering set of models are built by parsing the observations using action grammars that are most similar to ID\LP Grammars [39]. ID\LP grammars admit partial ordering, and Geib and Goldman further modify the parsing algorithm to allow multiple interleaved plans. The probabilities for these models are computed based on a Bayesian model of plan execution. This allows their system to handle multiple, interleaved, partially ordered plans as well as the failure to observe actions. They have

also proposed extensions to address partial observability and recognizing goal abandonment. This approaches' most significant limitation may be its need to maintain the covering set of explanations for a given set of observations. In some settings the cost of this process can be prohibitive.

Avrahami-Zilberbrand and Kaminka [6] have reported a approach similar to that of Geib and Goldman [44, 42, 41]. It differers in that they check the consistency of observed actions against previous hypotheses rather than using an action grammar for filtering possible explanations. This allows them to solve many of the same problems as addressed by Geib and Goldman but does reintroduced the problem of inference on the basis of failure to observe actions.

Hierarchal Hidden Markov Models (HHMMs) promise many of the efficiency advantages of parsing approaches, but with the additional advantages of supporting machine learning to automatically acquire their plan models. The first work that we know of in this area was provided by Bui [17] who has proposed a model of plan recognition based on a variant of HMMs. Unfortunately, in order to address multiple interleaved goals Bui, like Pynadath and Wellman, faces the problem of defining a probability distribution over the set of all possible root goal sets.

There's also work on cognitive assistive systems for the elderly by Liao, Fox, and Kautz [65] that makes use of HMMs. They use HMMs primarily to track the movements of their subjects, but incorporate information about possible routine movements through layered HMMs. The relative ease with which spatial regions can be decomposed and the consistent and simple transition probabilities between regions makes these problems very amenable to HMMs. When the application moves from these kinds of geographic domains to more symbolic domains as in computer network security the transition probabilities between states are much less clear and much harder to produce.

6 Towards a Unified Interpretation of Action

In order to conclude this review, we attempt to outline the overall "picture" of the action understanding and learning problem, and we pinpoint the sub-parts that are approached by the different communities and discuss their relationship and role for the action interpretation.

In order to investigate the full complexity of action we need to consider the following problem areas:

1. How to observe other agents: This concerns the detection, representation, recognition and interpretation of visually perceived actions of observed agents. Problems such as view-invariance, use of action grammars, pattern matching over time, representational issues, etc. need to be investigated. These are the problems presently investigated in the vision community.

2. How to control the physical body of a robot: this concerns learning/estimation of the mapping between the human and the robot kinematic chains.

3. How a robot can imitate other agents: this concerns how a robot can generalize over a small set of observed actions in order to generate novel ones from those observed. The two latter points are of present interest to the robotics community.

4. To arrive at a set of objects and a set of affordances for each of the objects, e.g., through observing, is a major interest in the AI community.

5. To arrive at a set of object-action complexes that take into account the acting agent when collecting sets of objects and object affordances.

6.1 Ego-Centric Action

In the robotics community recognition of human activity has been used extensively for robot task learning through imitation and demonstration [63, 101, 9, 79, 80, 66, 54, 30, 18]. Here, mainly human body model-based approaches (Sec. 3.3) are used. As mentioned above, one of the fundamentals of social behaviors of humans is the understanding of each others intentions through perception and recognition of performed actions. This is also underlined by the recent discovery of *mirror neurons* in the monkey's brain [87, 36]. The mirror neurons allow the monkey to interpret other's actions by aligning inside its mind the pose of its own (imagined) body to the pose of an observed one and appear to be of major importance for the ability of the monkey (and human) to learn through imitating others. Thus, the mirror neurons are a biological justification for the use of human body model-based approaches to recognizing actions.

By internally aligning the own body to an observed one, the mirror neurons move the reference system from the observed agent into the observer's ego-centric frame of reference. In imitation learning, the action to be learned is executed by the trainer in his/her own coordinate system. In other words, the robot observes the action in the trainers coordinate system and then, when imitating, recognizes and executes the observed action in its own coordinate system. The body model is often represented as a kinematic chain and the recognition is done in the space of possible joint configurations or Cartesian trajectories.

This *ego-centric* approach is in theory a great simplification of the action recognition task as one is able to compare and match the body movements of observed agents within a common, ego-centric, representation coordinate system.

The problems of the ego-centric approach are often due to the vision problem, i.e., the extraction of the visual data. As mentioned above, the quality of the visual data has to be sufficiently good and the tracked agent has to be large enough (in terms of pixels in height). First experiments [81, 107, 27] have been done in aiding the body tracking approaches with models for the executed action in order to constrain the tracking process. However, the models used so far are very simple and model usually periodic movements like walking. It is an open question and subject of present research how to incorporate more complex models to aid the tracking process.

Another problem stems from the general variability of even the simplest actions. Especially in every day like actions, simple movements such as "reach and grasp an object" can have different directions and reaching distances. To represent such actions, it is not sufficient to store simple trajectories. Instead, special care has to be taken that actions with different parameterizations can be recognized and synthesized. E.g., for the object grasping example, the action would be parameterized by the position of the object. One solution for parameterizing action from an ego-centric point of view was suggested by [127]. In their work, Wilson and Bobick model only simple actions, and it is not clear how this representation would scale to more complex actions.

6.2 Eco-Centric Action

For many actions that are meant to lead to a specific change in the environment, the precise way of how a teacher executes an action does sometimes not matter. Often, it cannot even be exactly repeated if, e.g., the object at which the action is aimed, is located at different positions.

Alternatively, a specific action may be carried out without any constraints on *how* it may be executed. The two examples from Sec. 1 on how to set-up and clean a dinner table are typical examples in this context: They are meant to cause a specific environmental change while the actual execution is either not particularly constrained or has to be planned on-line, depending on the present state of the environment. An observer can recognize the performed action by interpreting the change of the environment, e.g., "the table is set-up", without considering how the agent's actions that lead to the environmental change were precisely executed. This

viewpoint leads to an *eco-centric* interpretation of action as it puts the environment into the center of the action interpretation problem.

In order to approach this viewpoint one needs to consider two issues:

1. how to *represent* the changes in the environment and
2. how to physically *cause* specific changes in the environment.

The first issue contains three subproblems: a) How to *visually* recognize the changes in the environment is discussed in Sec. 3.1 and 3.2. b) How to *interpret* the changes is a matter of plan recognition (see Sec. 5). c) How to combine these two: The vision approach deals with continuous data while the approaches in AI deal with discrete data. A few attempts were made to connect the two approaches [96, 100]. Early approaches in robotics suggest [50] that changes in the environment should be represented as changes in the surface relationships between the scene objects.

The second issue is concerned with the execution of meaningful robot movements that are meant to cause a specific change in the environment. Again, this issue has a number of subproblems: a) How to *execute* a simple meaningful action. This is a problem beyond simple motor control (Sec. 6.1) as the execution is based on the state of the environment, e.g., the position of the object to be grasped. b) How to *plan* the meaningful action to be executed by the robot. This is a problem which is inversely related to the point b), above. It requires a usually grammatical representation that describes the possible changes of the environment and the physical actions that can cause them [112, 111] (see also Sec. 5).

6.3 Object Action Complexes

To formalize the possible changes in the environment, grammatical production rules for objects, object states and object affordances (an affordance changes the state of an object) cat be used. E. g. a door can have the states {*open,closed*} and the affordance {*close door, open door*}.

In some cases, the objects and production rules are *a-priori* specified by an expert and the scene state is usually considered to be independent from the presence of the agent itself within the scene, i.e., the agent affects the scene state only through a set of specified actions. The fact that an agent might physically not be able to execute a particular action, e.g., because it might not be in the right position or it might be too weak, must be taken into account. The research on *motion planning* takes this into account, while in most cases it is assumed that the scene (environment) does not change while the agent performs the planned movement.

Another problem that arises from *a-priori* definition of object affordances is the problem of taking into account the physical properties of the robot. In order for a robot to interact successfully with an environment, the set of object affordances it takes into account for planning must necessarily reflect its physical abilities. Unless the programmer has a precise model of the physical robot body as well as for the scene objects and the entire scene available, the affordances need to be learned by the robot itself through exploration. This leads us to the concept of object-action complexes. In order to learn how valid and appropriate an action is, the robot needs eventually to try to execute it. This could be interpreted as “playing” or “discovering”. Similarly to humans, the learning process can be biased through imitation learning as long as there is sufficient similarity between the learning agent and the teacher.

6.4 Outlook

To approach research in action at its full complexity by letting a robot system acquire its own experience and knowledge about movements, objects and possible world changes (and thus their interpretation) appears difficult at present. One possibility to limit the learning complexity is to

constrain experimental scenarios. Another possibility is to use *a-prior* knowledge at a suitable abstraction level.

Action understanding straddles in the gray zone between robotics, computer vision and AI and it has become a major thrust in robotics and computer vision. Unlike object recognition, which also plays a major role in this context and which was made tractable using specific geometric models of a physical object, the understanding of action requires reasoning about qualitative temporal relationships. Considerable research will be necessary to fully understand the problems associated with action understanding.

References

[1] J. Aggarwal and S. Park. Human Motion: Modeling and Recognition of Actions and Interactions. In *Second International Symposium on 3D Data Processing, Visualization and Transmission*, Thessaloniki, Greece, September 6-9 2004. 3, 4

[2] M. Ahmad and S. Lee. Human Action Recognition Using Multi-view Image Sequence Features. In *International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, April 10-12, 2006. 6

[3] M. A. Arbib. Perceptual structures and distributed motor control. In V. B. Brooks, editor, *Handbook of Physiology, Section 2: The Nervous System (Vol. II, Motor Control, Part 1)*, pages 1449–1480. American Physiological Society, 1981. 10

[4] T. Asfour, F. Gyarfas, P. Azad, and R. Dillmann. Imitation learning of dual-arm manipulation tasks in humanoid robots. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, Genoa, Italy, 2006. 11

[5] C. G. Atkeson and S. Schaal. Robot learning from demonstration. In *Proc. 14th International Conference on Machine Learning*, pages 12–20. Morgan Kaufmann, 1997. 11

[6] D. Avrahami-Zilberbrand and G. A. Kaminka. Fast and complete symbolic plan recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005. 14

[7] P. Azad, A. Ude, R. Dillmann, and G. Cheng. A full body human motion capture system using particle filtering and on-the-fly edge detection. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, Los Angeles, California, November 2004. 7

[8] D. C. Bentivegna, C. G. Atkeson, A. Ude, and G. Cheng. Learning to act from observation and practice. *International Journal of Humanoid Robotics*, 1(4):585–611, 2004. 10

[9] A. Billard. Imitation: A review. *Handbook of brain thory and neural network, M. Arbib (ed.)*, pages 566–569, 2002. 15

[10] A. Billard, Y. Epars, S. Calinon, S. Schaal, and G. Cheng. Discovering Optimal Imitation Strategies. *Robotics and Autonomous Systems*, 47:69–77, 2004. 10, 11

[11] B. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15-21, 2005. 6

[12] A. Bobick. Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion. *Philosophical Trans. Royal Soc. London*, 352:1257–1265, 1997. 3, 6

[13] A. Bobick and J. Davis. The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 6

[14] O. Boiman and M. Irani. Detecting Irregularities in Images and in Video. In *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15-21, 2005. 4

[15] G. Bradski and J. Davis. Motion Segmentation and Pose Recognition with Motion History Gradients. *Machine Vision and Applications*, 13(3):174–184, 2002. 6

[16] C. Breazeal and B. Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481–487, 2002. 9

[17] H. H. Bui, S. Venkatesh, and G. West. Policy recognition in the Abstract Hidden Markov Model. *Journal of AI Research*, 17:451–499, 2002. 14

[18] S. Calinon, F. Guenter, and A. Billard. Goal-Directed Imitation in a Humanoid Robot. In *International Conference on Robotics and Automation*, Barcelona, Spain, April 18-22, 2005. 10, 15

[19] E. Charniak and R. P. Goldman. A bayesian model of plan recognition. *Artificial Intelligence*, 64(1):53–79, 1993. 13

[20] E. Charniak and D. McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA, 1987. 13

[21] A. Chowdhury and R. Chellappa. A Factorization Approach for Activity Recognition. In *Computer Vision and Pattern Recognition*, Washington DC, USA, June, 2004. 5

[22] P. R. Cohen, C. R. Perrault, and J. F. Allen. Beyond question answering. In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Processing*, pages 245–274. Lawrence Erlbaum Associates, 1981. 12

[23] B. Dariush. Human Motion Analysis for Biomechanics and Biomedicine. *Machine Vision and Applications*, 14:202–205, 2003. 2

[24] J. Davis and H. Gao. Recognizing Human Action Efforts: An Adaptive Three-Mode PCA Framework. In *Internatinal Conference on Computer Vision*, Nice, France, Oct 13-16, 2003. 7

[25] J. Davis and H. Gao. Gender Recognition from Walking Movements using Adaptive Three-Mode PCA. In *Computer Vision and Pattern Recognition*, Washington DC, USA, June, 2004. 7

[26] J. Davis and S. Taylor. Analysis and Recognition of Walking Movements. In *International Conference on Pattern Recognition*, Quebec, Canada, August 11-15 2002. 7

[27] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, June 13-15 2000. 7, 15

[28] R. Dillmann. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems*, 47:109–116, 2004. 9

[29] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *Internatinal Conference on Computer Vision*, Nice, France, Oct 13-16, 2003. 5

[30] S. Ekvall and D. Kragic. Grasp recognition for programming by demonstration tasks. In *IEEE International Conference on Robotics and Automation, ICRA'05*, pages 748 – 753, 2005. 15

[31] S. Ekvall and D. Kragic. Learning task models from multiple human demonstrations. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN'06*, pages 358–363,, 2006. 12

[32] A. Elgammal and C. Lee. Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In *Computer Vision and Pattern Recognition*, Washington DC, USA, June, 2004. 6

[33] A. Elgammal and C. Lee. Separating Style and Content on a Nonlinear Manifold. In *Computer Vision and Pattern Recognition*, Washington DC, USA, June, 2004. 6

[34] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis. Learning Dynamics for Exemplar-based Gesture Recognition. In *Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, June 16-22, 2003. 7

[35] D. N. et al. The objective basis of behavior unit. *Journal of Personality and Social Pshychology*, 35(12):847–862, 1977. 9

[36] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Visuomotor neurons: Ambiguity of the discharge or 'motor perception'? *International Journal of Psychophysiology*, 35(2-3):165–177, 2000. 15

[37] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid Models for Human Motion Recognition. In *Computer Vision and Pattern Recognition*, San Diego, California, USA, June 20-25, 2005. 7, 8

[38] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning About Objects Through Action - Initial Steps Towards Artificial Cognition. In *IEEE International Conference on Robotics and Automation*, pages 3140–3145, 2003. 9

[39] J. G. Edward Barton. On the complexity of ID/LP parsing. *Computational Linguistics*, 11(4):205–218, 1985. 13

[40] J. Gao, A. Hauptmann, and H. Wactlar. Combining Motion Segmentation with Tracking for Activity Analysis. In *International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, May 17-19 2004. 6

[41] C. Geib. Plan recognition. In A. Kott and W. McEneaney, editors, *Adversarial Reasoning*. Chapman and Hall/CRC, 2007. 13, 14

[42] C. W. Geib and R. P. Goldman. Recognizing plan/goal abandonment. In *Proc. Int. Joint Conference on Artificial Intelligence*, 2003. 13, 14

[43] M. Giese and T. Poggio. Neural Mechanisms for the Recognition of Biological Movements. *Nature Reviews*, 4:179–192, 2003. 2

[44] R. P. Goldman, C. W. Geib, and C. A. Miller. A new model of plan recognition. In *Proceedings of the 1999 Conference on Uncertainty in Artificial Intelligence*, 1999. 13, 14

[45] J. Gonzàlez, J. Varona, F. Roca, and J. Villanueva. *aSpaces*: Action Spaces for Recognition and Synthesis of Human Actions. In *International Workshop on Articulated Motion and Deformable Objects*, Palma de Mallorca, Spain, Nov 21-23, 2002. 3

[46] L. Gorelick, M. Galun, E. Sharon, A. Brandt, and R. Basri. Shape Representation and Recognition Using the Poisson Equation. In *Computer Vision and Pattern Recognition*, Washington DC, USA, June, 2003. 6

[47] A. Gritai, Y. Sheikh, and M. Shah. On the Use of Anthropometry in the Invariant Analysis of Human Actions. In *International Conference on Pattern Recognition*, Cambridge, UK, Aug 23-26, 2004. 7

[48] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-Time Surveillance of People and Their Activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000. 7

[49] A. Ijspeert, J. Nakanishi, and S. Schaal. Movement Imitation withNonlinear Dynamical Systems in Humanoid Robots. In *International Conference on Robotics and Automation*, Washington DC, USA, May, 2002. 10

[50] K. Ikeuchi and T. Suehiro. Towards assembly plan from observation, Part I: Task recognition with polyhedral objects. *IEEE Transactions on Robotics and Automation*, 10(3):368–385, 1994. 11, 16

[51] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, 23(4-5):363–377, 2004. 10, 11

[52] Y. Ivanov and A. Bobick. Recognition of Visual Activities and Interactions by Stochastic Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000. 8

[53] O. Jenkins and M. Mataric. Automated Modularization of Human Motion into Actions and Behaviors. Technical Report CRES-02-002, Center for Robotics and Embedded Systems, University of S. California, 2002. 3

[54] O. C. Jenkins and M. J. Mataric. Performance-derived behavior vocabularies: Data-driven acqusition of skills from motion. *International Journal of Humanoid Robotics*, 1(2):237–288, Jun 2004. 10, 15

[55] H. Jiang, M. Drew, and Z. Li. Successive convex matching for action detection. In *Computer Vision and Pattern Recognition*, pages 1646–1653, New York City, New York, USA, June 17-22, 2006. 6

[56] S. B. Kang and K. Ikeuchi. Toward automatic robotic instruction from perception – Temporal segmentation of tasks from human hand motion. *IEEE Transactions on Robotics and Automation*, 11(5):670–681, 1995. 11

[57] S. B. Kang and K. Ikeuchi. Toward automatic robotic instruction from perception – Mapping human grasps to manipulator grasps. *IEEE Transactions on Robotics and Automation*, 13(1):81–95, 1997. 11

[58] H. Kautz. *A Formal Theory of Plan Recognition and its Implementation*. PhD thesis, University of Rochester, 1991. 13

[59] H. Kautz and J. F. Allen. Generalized plan recognition. In *Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI-86)*, pages 32–38, 1986. 12

[60] D. Kendall, D. Barden, T. Carne, and H. Le. *Shape and Shape Theory*. Wiley, 1999. 5

[61] K.Ogawara, J.Takamatsu, K.Kimura, and K.Ikeuchi. Generation of a task model by intergrating multiple observations of human demonstrations. In *Proceedings of the IEEE Intl. Conf. on Robotics and Automation (ICRA '02)*, pages 1545–1550, May 2002. 12

[62] V. Krueger. Recognizing action primitives in complex actions using hidden markov models. In *Advances in Visual Computing*, pages 538–547, Second Int. Symp. on Visual Computing, Lake Tahoe, NV, USA, November 6-8, 2006. 8

[63] Y. Kuniyoshi, M. Inaba, and H. Inoue. Learning by watching, extracting reusable task knowledge from visual observation of human performance. In *IEEE Transactions on Robotics and Automation*, volume 10(6), pages 799–822, 1994. 11, 15

[64] Y. Kuniyoshi, Y. Yorozu, M. Inaba, and H. Inoue. From visuo-motor self learning to early imitation - a neural architecture for humanoid learning. In *IEEE International Conference on Robotics and Automation*, pages 3132–3139, Taipei, Taiwan, 2003. 11

[65] L. Liao, D. Fox, and H. A. Kautz. Location-based activity recognition using relational Markov networks. In *Proc. Int. Joint Conference on Artificial Intelligence*, pages 773–778, 2005. 14

[66] M. C. Lopes and J. santos Victor. Visual transformations in gesture imitation: What you see is what you do. In *IEEE International Conference on Robotics and Automation, ICRA04*, pages 2375– 2381, 2003. 15

[67] C. Lu and N. Ferrier. Repetitive Motion Analysis: Segmentation and Event Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):258–263, 2004. 8

[68] Y. Luo, T.-W. Wu, and J.-N. Hwang. Object-Based Analysis and Interpretation of Human Motion in Sports Video Sequences by Dynamic Bayesian Networks. *Computer Vision and Image Understanding*, 92:196–216, 2003. 7

[69] F. Lv and R. Nevatia. Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost. In *European Conference on Computer Vision*, Graz, Austria, May 7-13, 2006. 8

[70] M. J. Matarić, M. Williamson, J. Demiris, and A. Mohan. Behavior-based primitives for articulated control. In *Proc. Fifth Int. Conf. on the Simulation of Adaptive Behavior*, pages 165–170, Cambridge, Mass., 1998. MIT Press. 10

[71] J. McCarthy. Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39,171–172, 1980. 13

[72] A. N. Meltzoff and M. K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78, 1977. 11

[73] H. Miyamoto, S. Schaal, F. Gandolfo, Y. Koike, R. Osu, E. Nakano, Y. Wada, and M. Kawato. A kendama learning robot based on bi-directional theory. *Neural Networks*, 9:1281–1302, 1996. 11

[74] T. Moeslund, A. Hilton, and V. Krueger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–127, 2006. 4

[75] F. A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Phil. Trans. R. Soc. Lond. B*, 355:1755–1769, 2000. 10

[76] H.-H. Nagel. From Image Sequences Towards Conceptual Descriptions. *Image and Vision Computing*, 6(2):59–74, 1988. 3

[77] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato. Learning from demonstration and adaptation of biped locomotion. *Robotics and Autonomous Systems*, 47:79–91, 2004. 11

[78] C. Nehaniv and K. Dautenhahn. Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications. In J. Demiris and A. Birk, editors, *Learning Robots: An Interdisciplinary Approach*. World Scientific Press, 1999. 10

[79] K. Ogawara, S. Iba, H. Kimura, and K. Ikeuchi. Recognition of human task by attention point analysis. In *IEEE International Conference on Intelligent Robot and Systems IROS'00*, pages 2121–2126, 2000. 11, 15

[80] K. Ogawara, S. Iba, H. Kimura, and K. Ikeuchi. Acquiring hand-action models by attention point analysis. In *IEEE International Conference on Robotics and Automation*, pages 465–470, 2001. 15

[81] D. Ormoneit, H. Sidenbladh, M. Black, and T. Hastie. Learning and Tracking Cyclic Human Motion. In *Workshop on Human Modeling, Analysis and Synthesis at CVPR*, Hilton Head Island, South Carolina, June 13-15 2000. 7, 15

[82] R. W. Paine and J. Tani. Motor primitive and sequence self-organization in a hierarchical recurrent neural network. *Neural Networks*, 17(8-9):1291–1309, 2004. 10

[83] V. Parameswaran and R. Chellappa. View Invariance for Human Action Recognition. *International Journal of Computer Vision*, 66(1):83–101, 2006. 7

[84] S. Park and J. Aggarwal. Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy. In *CVPR workshop on Articulated and non-rigid motion*, Washington DC, USA, June, 2004. 8

[85] D. Pynadath and M. Wellman. Probabilistic state-dependent grammars for plan recognition. In *Proc. Int. Conf. on Uncertainty in AI*, pages 507–514, 2000. 13

[86] M. Rahman and A. Robles-Kelly. A Tuned Eigenspace Technique for Articulated Motion Recognition. In *European Conference on Computer Vision*, Graz, Austria, May 7-13, 2006. 6

[87] V. Ramachandran. Mirror neurons and imitation learning as the driving force behind the gerat leap forward in human evolution. *Edge*, 69, 2000. 15

[88] C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. *Journal of Computer Vision*, 50(2):203–226, 2002. 8

[89] H. Ren and G. Xu. Human Action Recognition with Primitive-based Coupled-HMM. In *International Conference on Pattern Recognition*, Quebec, Canada, August 11-15 2002. 7

[90] H. Ren, G. Xu, and S. Kee. Subject-independent Natural Action Recognition. In *International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, May 17-19, 2004. 7

[91] Y. Ricquebourg and P. Bouthemy. Real-Time Tracking of Moving Persons by Exploiting Spatio-Temporal Image Slices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):797–808, 2000. 6

[92] M. Riley, A. Ude, and C. G. Atkeson. Methods for motion generation and interaction with a humanoid robot: Case studies of dancing and catching. In *AAAI and CMU Workshop on Interactive Robotics and Entertainment*, pages 35–42, Pittsburgh, PA, April 2000. 11

[93] J. Rittscher, A. Blake, and S. Roberts. Towards the Automatic Analysis of Complex Human Body Motions. *Image and Vision Computing*, 20:905–916, 2002. 6

[94] G. Rizzolatti, L. Fogassi, and V. Gallese. Parietal Cortex: from Sight to Action. *Current Opinion in Neurobiology*, 7:562–567, 1997. 2

[95] G. Rizzolatti, L. Fogassi, and V. Gallese. Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action. *Nature Reviews*, 2:661–670, Sept. 2001. 2

[96] N. Robertson and I. Reid. Behaviour Understanding in Video: A Combined Method. In *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15-21, 2005. 5, 16

[97] M. Roh, B. Christmas, J. Kittler, and S. Lee. Robust Player Gesture Spotting and Recognition in Low-Resolution Sports Video. In *European Conference on Computer Vision*, Graz, Austria, May 7-13, 2006. 5

[98] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *SCIENCE*, 290:2323–2327, 2000. 6

[99] M. Ruchanurucks, S. Nakaoka, S. Kudo, and K. Ikeuchi. Humanoid robot motion generation with sequential physical constraints. In *IEEE International Conference on Robotics and Automation*, pages 2649–2654, Orlando, Florida, 2006. 11

[100] K. Sato and J. Aggarwal. Tracking and Recognizing Two-person Interactions in Outdoor Image Sequences. In *Workshop on Multi-Object Tracking*, Vancouver, Canada, July 8 2001. 6, 16

[101] S. Schaal. Is Imitation Learning the Route to Humanoid Robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999. 2, 9, 15

[102] S. Schaal. Dynamic movement primitives – A framework for motor control in humans and humanoid robotics. In *Proc. International Symposium on Adaptive Motion of Animals and Machines*, pages 12–20, 2003. 10

[103] C. Schmidt, N. Sridharan, and J. Goodson. The Plan recognition problem: an intersection of psychology and artificial intelligence. *Artificial Intelligence*, 11:45–83, 1978. 12

[104] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978. 5

[105] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the Space of Human Action. In *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15-21, 2005. 7

[106] Y. Shi, A. Bobick, and I. Essa. Learning Temporal Sequence Model from Partially Labeled Data. In *Computer Vision and Pattern Recognition*, New York City, New York, USA, June 17-22, 2006. 7

[107] H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, Copenhagen, Denmark, 2002. 7, 15

[108] C. L. Sidner. Plan parsing for intended response recognition in discourse. *Computational Intelligence*, 1(1):1–10, 1985. 13

[109] C. Sminchiescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *International Conference on Computer Vision, ICCV'05*, pages 1808–1815, 2005. 11

[110] C. Stauffer and W. Grimson. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000. 4

[111] M. Steedman. Temporality. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 895–938. Elsevier, 1997. 16

[112] M. Steedman. Plans, affordances, and combinatory grammar. *Linustics and Philosophy*, 25:723–753, 2002. 16

[113] A. Stolcke. An Efficient Probabilistic Context-Free Parsing Algorithm That Computes Prefix Probabilities. *Computational Linguistics*, 21(2):165–201, 1995. 8

[114] A. Stoytchev. Behavior-Grounded Representation of Tool Affordances. In *IEEE International Conference on Robotics and Automation*, pages 3060–3065, 2005. 9

[115] J. Tenenbaum, V. de Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *SCIENCE*, 290:2319–2323, 2000. 6

[116] C. Tomasi and T. Kanade. Shape and Motion from Image Streams under Orthography: A Factorization Method. *International Journal of Computer Vision*, 9:137–154, 1992. 5

[117] A. Ude. Robust estimation of human body kinematics from video. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1489–1494, Kyongju, Korea, 1999. 11

[118] A. Ude, C. G. Atkeson, and M. Riley. Programming full-body movements for humanoid robots by observation. *Robotics and Autonomous Systems*, 47:93–108, 2004. 11

[119] N. Vasvani, A. R. Chowdhury, and R. Chellappa. Activity Recognition Using the Dynamics of the Configuration of Interacting Objects. In *Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, June 16-22, 2003. 5

[120] D. Vecchio, R. Murray, and P. Perona. Decomposition of Human Motion into Dynamics-based Primitives with Application to Drawing Tasks. *Automatica*, 39(12):2085–2098, 2003. 8

[121] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury. The Function Space of an Activity. In *Computer Vision and Pattern Recognition*, New York City, New York, USA, June 17-22, 2006. 4

[122] M. Vilain. Deduction as parsing. In *Proceedings of the Conference of the American Association of Artifical Intelligence (1991)*, pages 464–470, 1991. 13

[123] J. Wang and B. Bodenheimer. An Evaluation of a Cost Metric for Selecting Transitions between Motion Segments. In *SIGGRAPH Symposium on Computer Animation*, 2003. 8

[124] L. Wang, H. Ning, T. Tan, and W. Hu. Fusion of Static and Dynamic Body Biometrics for Gait Recognition. In *International Conference on Computer Vision*, Nice, France, 13-16 October 2003. 7

[125] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003. 7

[126] R. Wilensky. *Planning and Understanding*. Addison-Wesley, 1983. 12

[127] A. Wilson and A. Bobick. Parametric hidden markov models for gesture reocngition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:884–900, September 1999. 15

[128] T. Xiang and S. Gong. Beyond Tracking: Modelling Action and Understanding Behavior. *International Journal of Computer Vision*, 67(1):21–51, 2006. 5

[129] M. Yamamoto, H. Mitomi, F. Fujiwara, and T. Sato. Bayesian Classification of Task-Oriented Actions Based on Stochastic Context-Free Grammar. In *International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, April 10-12, 2006. 8

[130] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 379–385, June 1992. 6

[131] A. Yilmaz and M. Shah. Actions Sketch: A Novel Action Representation. In *Computer Vision and Pattern Recognition*, San Diego, California, USA, June 20-25, 2005. 6

[132] A. Yilmaz and M. Shah. Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras. In *Internatinal Conference on Computer Vision*, Beijing, China, Oct 15-21, 2005. 7

[133] H. Yu, G.-M. Sun, W.-X. Song, and X. Li. Human Motion Recognition Based on Neural Networks. In *International Conference on Communications, Circuits and Systems*, Hong Kong, China, May 2005, 2005. 6