Advanced Robotics, Vol. 21, No. 13, pp. 1473–1501 (2007)

Also available online - www.brill.nl/ar

Review

The meaning of action: a review on action recognition and mapping

VOLKER KRÜGER [1,*], DANICA KRAGIC [2], ALEŠ UDE [3] and CHRISTOPHER GEIB [4]

[1] *Computer Vision and Machine Intelligence Laboratory, Aalborg University, DK-3750 Ballerup, Denmark*
[2] *Computer Vision and Active Perception Laboratory, CSC-KTH, SE-10044 Stockholm, Sweden*
[3] *Jozef Stefan Institute, Department of Automatics, Biocybernetics & Robotics, 1000 Ljubljana, Slovenia*
[4] *School of Informatics, University of Edinburgh, Edinburgh, EH8 9LE UK*

Received 12 December 2006; revised 2 May 2007; accepted 16 May 2007

Abstract—In this paper, we analyze the different approaches taken to date within the computer vision, robotics and artificial intelligence communities for the representation, recognition, synthesis and understanding of action. We deal with action at different levels of complexity and provide the reader with the necessary related literature references. We put the literature references further into context and outline a possible interpretation of action by taking into account the different aspects of action recognition, action synthesis and task-level planning.

Keywords: Action recognition; action representation; action synthesis; action understanding; planning.

1. INTRODUCTION

The recognition and interpretation of human- or robot-induced actions and activities has gained considerable interest in the computer vision, robotics and artificial intelligence communities. This is partially due to increasing computer power that allows large amounts 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, robot learning and control. Depending on the application, starting points and aims in action-related research are different. In this paper, we analyze the different approaches to action representation, recognition and mapping taken to date within the three communities.

*To whom correspondence should be addressed. E-mail: vok@cvmi.aau.dk

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 potential is in automatic scene understanding systems that include 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 one 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, e.g., as 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 strong attention from the computer vision community. Here, action recognition is often treated as a pattern-matching problem with an additional time dimension. Attention is given to improper imaging conditions, noisy input data, and the development of robust approaches for representation and recognition the actions.

There is strong neurobiological evidence that human actions and activities are directly connected to the motor control of the human body [1–3]. 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 [4–6] where the goal of imitation learning is to develop robot systems that are able to relate perceived actions of another (human) agent to their own embodiment in order to learn and later to recognize and to perform the demonstrated actions. One of the goals for the future is to enable artifical agents to acquire novel behaviors through observation of humans or other agents.

The neurobiological findings motivate research to identify a set of action primitives that allow (i) representation of the visually perceived action and (ii) 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 greatly 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:

(i) 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.

(ii) 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.

(iii) Given a specific scene state, the robot may be unable to perform a particular action. For example, the representation may specify that wine glasses can be piled on top of plates, but the robot may be unable to reach the desired height.

(iv) The entire scene may change during the execution phase, and the robot has to be able to react to sudden changes and replan its task.

Different aspects of the above problems have been considered in the area of task planning and sequencing with the specific focus on structured collections of actions. Here, different types of reasoning systems have been proposed, including rule-based systems, traditional Bayes nets, context free grammars, etc., mainly for task planning purposes. Different methods and levels of action representation make the strongest obstacle to integrating the requirements for high-level conceptual state change representations suitable for planning and low-level continuous action execution and imitation for robots.

In spite of the differences in the potential applications, 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:

(i) Recognize the movements and actions of observed agents (recognizing the action by observing it).

(ii) Understand what effects certain actions have on the environment of the actor (recognizing the action by observing its effects on the environment).

(iii) Understand how to physically perform a certain action in order to cause a particular change in the environment.

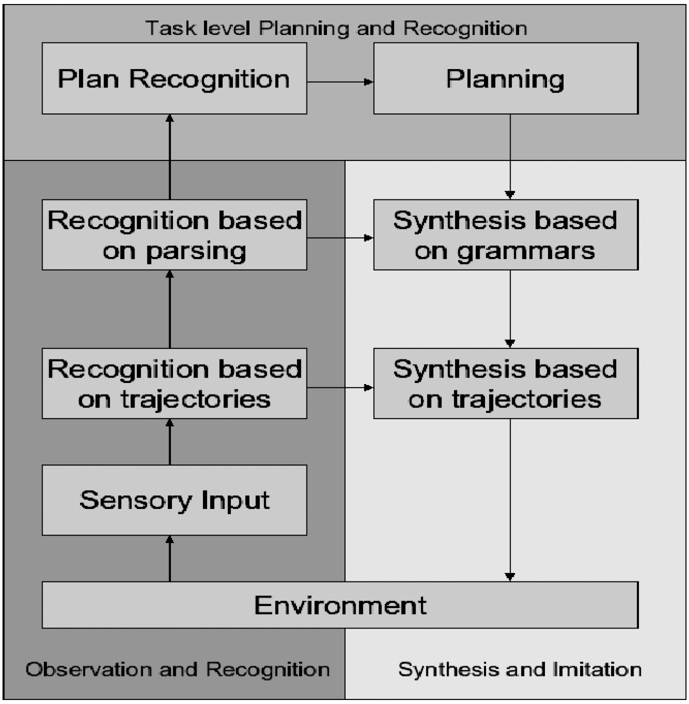

Figure 1. Different levels of action consideration discussed in this paper: Observation and recognition (bottom left, Section 3), synthesis and imitation (bottom right, Section 4), and task-level planning and recognition (top, Section 5).

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 capabilities.

In this paper, we analyze the different approaches taken to date for dealing with action at different levels of complexity (see Fig. 1) and provide the reader with the necessary related literature references. Different authors use different terms for discussing action primitives and action grammars. In Section 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 Sections 3–5 we discuss how the representation and recognition of actions is treated with respect to representation, synthesis and planning. We conclude this paper in Section 6.

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 [7], Nagel suggested to use a hierarchy of change, event, verb, episode, history. An alternative hierarchy (reflecting the computational aspects) is proposed by Bobick [8] who suggests to use movement, activity and action as different levels of abstraction (see also Ref. [9]). Others suggest to also include situations [10] or use a hierarchy of action primitives and parent behaviors [11].

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 Section 3.4).

As an example, in tennis, action primitives could be '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.

Good overviews of activity recognition are given by Aggarwal *et al.* [9, 12], in Refs [13, 14], as well as in a more recent one by Moeslund *et al.* [15]. 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.* [16] discuss the structure of an action and activity space.

2.1. Outlook

In order to investigate the full complexity of systems that deal with action representation, recognition and synthesis we need to consider the following problem areas (see Table 1 for a summary):

(i) 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.

(ii) How to control the physical body of a robot. This concerns learning/estimation of the mapping between the human and the robot kinematic chains.

(iii) How a robot can imitate other agents. This concerns how a robot can generalize over a set of observed actions in order to generate novel ones from those observed. Here, issues such as hierarchical organization for sequences and

Table 1.
Researach areas, input and issues to investigate

Problem	Input	Issues to investigate
Recognition	Two-dimensional image data, sensor data, MoCap data	View variance, variance in execution direction/scale, learning and using grammars
Imitation	Set of learned actions, object information (position, type, shape, orientation, etc.), robot kinematics, human body model	Object-dependent execution of action, mapping between kinematic chains, generalization over different observations, planning
Learning object affordances	Agent, scene with known objects	Learning statistics between objects, scene state changes and agent's actions through observation and self-exploration

probabilistic inference and planning for recognition, prediction and decision making are relevant.

(iv) Learning objects and their affordances, thus arriving at a set of object–action complexes that take into account the acting agent and the context.

2.2. Ego-centric action

In the robotics community, recognition of human activity has been used extensively for robot task learning through imitation and demonstration [5, 17–24]. Here, mainly human body model-based approaches (Section 3.3) are used. 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 mirror neurons, motor resonance or mirroring [25–29]. The mirror neurons allow the monkey to interpret others' 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. The quality of the visual data has to be sufficiently good and the tracked agent has to be large enough (in terms of pixels). First experiments [30–32] 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 constrain 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 Ref. [33]. However, in this work only simple actions are modeled and it is not clear how this representation would scale to more complex actions.

The work presented in Refs [21, 34, 35] proposes a general architecture for action (mimicking) and program (gesture)-level visual imitation. The authors present a holistic approach to the problem by facilitating (i) the use of motor information for gesture recognition, (ii) usage of context (e.g., object affordances) to focus the attention of the recognition system and reduce ambiguities, and (iii) the use of iconic image representations for the hand, as opposed to fitting kinematic models to the video sequence.

2.3. 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 sometimes does not matter. Often, it cannot even be exactly repeated if, for example, 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 Section 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:

(i) How to represent the changes in the environment.

(ii) 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 Sections 3.1 and 3.2. (b) How to interpret the changes is a matter of plan recognition (see Section 5). (c) How to combine these two: a few attempts were made to connect the two approaches [36, 37]. Some of the early approaches in robotics suggest [38] 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 (Section 2.2) 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 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 [39, 40].

2.4. 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) can be used [41, 42], 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’. Similar 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.

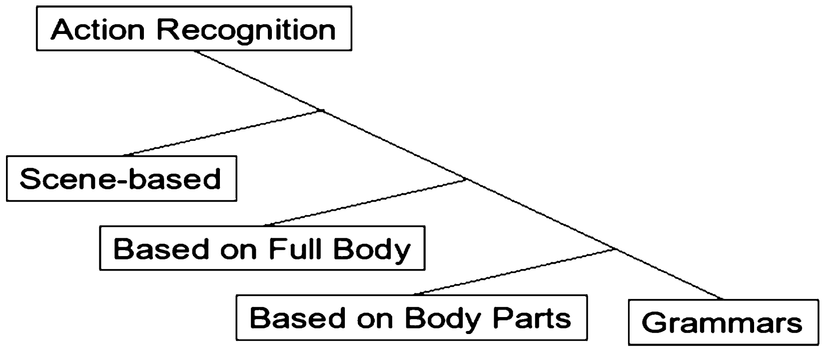

Figure 2. Different types of action recognition approaches presented in Section 3.

3. INTERPRETATION AND RECOGNITION OF ACTION

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 is usually not meant to distinguish explicitly, e.g., cars from humans. On the other hand, some action recognition applications focus explicitly on human activities and the interactions between human agents. Here, we follow Aggarwal and Cai [12] and distinguish between the full-body-based approaches that model the human either as a whole, i.e., without distinguishing between body parts, and the body-part-based approaches that model the human in a detailed manner as a set of body parts. Most full-body approaches attempt to identify information such as gender, identity, or simple actions like walking or running. Researchers using human body-part-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. Body-part-based approaches can also be used in medical applications or in applications from the entertainment industry. In the following, we review some of the recent publications that have their emphasis on action recognition. The presented techniques use video camera data as their primary input source, use mostly well-known tracking and motion capture techniques, and discuss how the results of those techniques can be used for action recognition. For papers on vision-based motion capture and tracking, we refer the reader to Ref. [15]. Figure 2 summarizes the different types of approaches we have mentioned in this section.

3.1. Scene interpretation

Many approaches consider the camera view as a whole, and attempt to learn and recognize activities by observing the motion of objects without necessarily knowing their identity, i.e., by identifying the changes in the scene over time. This is reasonable in situations where the objects are small enough to be represented as points on a two-dimensional (2-D) plane.

Stauffer *et al.* [43] 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 [44] approach the problem of detecting 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 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 Refs [45, 46] activity trajectories are modeled using non-rigid shapes and a dynamic model that characterizes the variations in the shape structure. Vaswani *et al.* [46] use Kendall's statistical shape theory [47]. Non-linear 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.* [45] use a subspace method to model activities as a linear combination of 3-D basis shapes. The work is based on the factorization theorem [48]. Deviations from the learned normal activity shapes can be used to identify abnormal ones. A similar complex task is approached by Xiang and Gong [49]. They present a unified bottom-up and top-down approach to model complex activities of multiple objects in cluttered scenes in order to (i) learn statistical dependencies between the objects, (ii) structure and parameters, (iii) select visual features that represent activities of multiple objects, (iv) infer semantic descriptions of activities from the learned model and (v) discuss how to use the activity model to improve interpretation of individual objects. Their approach is object-independent (it can be body parts, cars, etc.) and they use a dynamically multi-linked hidden Markov models (HMMs) to interlink between multiple temporal processes corresponding to multiple event classes.

3.2. Recognizing human actions without using body parts

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 2-D image projections, without the intermediate use, for example, of a 3-D human model. The argument is that the use of an explicit human (not necessarily 3-D) model is often not feasible in case of noisy and imperfect imaging conditions, and that a direct pattern recognition based on the 2-D data is potentially more robust. This argument holds especially when there are only very few pixels on the image of the observed agent.

A pioneering work has been presented by Efros *et al.* [50]. 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. The work of Robertson and Reid [36] extends the work of Efros [50] 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*.

A large number of publications work with space–time volumes, which is a recently proposed representation for the spatio-temporal domain. The 3-D contour of a person gives rise to a 2-D 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 [51, 52]. Articulated motions of a human then show a typical trajectory pattern. Ricquebourg and Bouthemy [51] demonstrate how XT-slices can facilitate tracking and reconstruction of 2-D motion trajectories. The reconstructed trajectory allows a simple classification between pedestrians and vehicles. Ritscher *et al.* [52] 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. Yilmaz and Shah [53] 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 3-D objects. Blank *et al.* [54] also analyze the XYT volume. They generalize techniques for the analysis of 2-D shapes [55] 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 [55]. 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 pattern recognition approaches such as PCA on sequences of silhouettes [56–58]. Bobick and Davis pioneered the idea of temporal templates [8, 59]. They propose a representation and recognition theory [8, 59] 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.* [60] 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. In Refs [61, 62], Elgammal and Lee use local linear embedding (LLE) [63, 64] 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 [62] as well as to infer 3-D body pose from 2-D silhouettes [61]. Sato and Aggarwal [37] 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. The classification is based on the computed feature vectors and the nearest mean classifier. In a number of publications, recognition is based on HMMs and dynamic Bayes networks (DBNs). The work of Yamato *et al.* [65] 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.* [66] 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.* [67] 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.* [68] 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 Section 3.2 about the difficulties in detecting individual 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. Ref. [69], start out with silhouettes and detect the body parts using a method inspired by the W4 system [70], 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 3-D model-based body-tracking approaches where the recognition of (periodic) action is used as a loop-back to support pose estimation [30–32, 71]. 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 [72–75]. Other approaches circumvent the vision problem by using a (vision-based) motion capture system in order to be able to focus on finding good representations of actions [76, 77].

Ren and Xu [78] 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 Ref. [79] by introducing primitive-based DBNs. One of the major obstacles in action recognition from images is the variability of the visual data under changing viewing directions. Parameswaran and Chellappa [77] consider the problem of view-invariant action recognition based on point-light displays by investigating 2-D and 3-D invariant theory. As no general, non-trivial 3-D–2-D invariants exist, Ref. [77] employ a convenient 2-D invariant representation by decomposing and combining the patches of a 3-D scene. For example, key poses can be identifies where joints in the different poses are aligned. In the 3-D case, six-tuples corresponding to six joints give rise to 3-D 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 [75] 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.* [72] address the invariant recognition of human actions and investigate the use of anthropometry to provide constraints on matching. They 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 3-D space. Sheikh *et al.* [73] pick up these results of Refs [72, 75], 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. Fanti *et al.* [74] present an approach that is content with a very small amount of user interaction for learning. They represent a human activity as a collection of body parts moving in a specific pattern. To find the most likely model alignment with input data they exploit appearance information which remains approximately invariant within the same setting. Then, they use expectation maximization (EM) 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 on the observed motion pattern.

3.4. Action primitives and grammars

Some of the work attempts 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 [80] employ techniques from the dynamic 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.* [81] 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.

While most scientists concentrate on the action representation by circumventing the vision problem, Rao *et al.* [82] 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 2-D 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 [83]. The system they describe has three 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 [84] 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 Ref. [84] they pick up a work by Stolcke [85] 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. Krüger [86] 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.* [87] 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 Ref. [88] where the authors are interested in recognizing and segmenting full-body human action. They decompose the large joint space into a set of feature spaces where each feature corresponds to a single joint or combinations of related joints. They then use HMMs to recognize each action class based on the features, and an AdaBoost scheme to detect and recognize the features.

4. ACTION LEARNING AND IMITATION

Unlike vision, robotics is mainly concerned with generative models of action that enable imitation learning. The robotics community has 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 [5, 89, 90]. We refer to Ref. [6] for an extensive overview of computational approaches to motor learning by imitation.

Hence, the interest of roboticist is to enable robots with action synthesis 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.

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 [41, 91]. 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. While most of the work on robotic grasping so far has dealt with analytical methods where the shape of the objects being grasped is known *a priori*, the goal for the future is to enable robots to learn how to manipulate novel objects independently and one way of bootstrapping the learning process may be through observation.

Some interesting questions arise:

- 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?

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 robot's representations about objects and the world. Furthermore, the successfully performed grasps can be used as ground truth for future grasp refinement [41].

4.1. Movement primitives

Many of the generative approaches have found their roots in the work of Newton *et al.* [92] 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 [93] 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 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 Ref. [94], 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 Section 3.4). Proposals include nonlinear dynamic attractor systems that can be flexibly adjusted to represent arbitrarily complex motor behaviors [95], primitive flow fields acquired from the motion capture data [22], hierarchical recurrent neural networks [96], HMMs [97, 98] and movement representation by force fields [99]. There may well be that no single representation exists and that different movement primitives are encoded differently.

More specifically, Jenkins *et al.* [22] suggest to apply a spatio-temporal nonlinear 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 and co-workers [95, 100] 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.* [97] use a 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.* [24] 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 [96] 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 [86]. 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 [101] or acquired from the motion capture data with the help of manual segmentation.

4.2. Imitation learning

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: (i) what to imitate, (ii) how to imitate and (iii) when to imitate [102]. 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 work concerned with the first two issues.

Robotics research on imitation started in the early 1990s under 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 [17, 38]. Kuniyoshi *et al.* [17] and Kang and Ikeuchi [103] 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 [104]. The mapping of whole-body human movements, e.g., dance movements, to the movements of a humanoid robot followed [105, 106]. An automatic approach to relate human kinematics to humanoid robot kinematics has been developed [107] and it has been shown how to incorporate balancing controllers into the captured movements [108].

Kuniyoshi *et al.* [109] 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 [110] 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.* [109] 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.* [111] 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 [112] 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.* [113] introduced a framework for the learning of walking controllers using dynamic movement primitives. Asfour *et al.* [114] use HMMs to generalize movements demonstrated to a robot several times.

However, a higher level of abstraction is achieved by sequencing a number of action units. HMMs have been proposed as a suitable representation for this purpose [19, 97, 98, 114]. 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 [115].

Billard *et al.* [97] 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.* [98] 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.

Herzog *et al.* [137] propose an exemplar-based approach for generating and recognizing arm movements. They use locally linear interpolated parametric HMMs. In their paper, the parameters for the HMMs are specified by the 2D position of an object on a table. Given the 2D parameters, the robot is able to grasp the object; given a grasping or pointing movement with an object, the system is able to identify the 2D position of the involved object on the table (i.e., the parameters of the parametric HMM).

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. 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.* [116], 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 [117] 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. PLAN AND INTENTION RECOGNITION

Task-level work in action and plan recognition has focused more on recognizing structured collections of actions and their interaction also in relation to development of cognitive architectures [42, 118]. 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 works in plan recognition [119, 120] were 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 [121] distinguishes between two kinds of plan recognition: intended and keyhole. In intended recognition, the agent is cooperative and its actions are done with the intent 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 [122] 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 [123].

Kautz also presented an approximate implementation of this approach that recasts the problem as one of computing vertex covers of the plan graph [124]. 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 [125] argued that plan recognition is just abduction, or reasoning to the best explanation [126], 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: (i) they can assert that the action has not and will not occur or (ii) they can be silent about whether an action has occurred—implying that the system has failed to notice the action, not that the action has not occurred. Both of these solutions are unsatisfying.

Both Vilain [127] and Sidner [128] 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 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 [129] have proposed probabilistic parsing for plan recognition. Using plans represented as probabilistic context-free grammars 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 might overcome this problem, but it is significantly more difficult to define a probability distribution for a probabilistic context-sensitive grammar.

Geib and Goldman [130–132] 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 [133]. 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 [134] have reported a approach similar to that of Geib and Goldman [130–132]. It differs 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 reintroduce the problem of inference on the basis of failure to observe actions.

Hierarchal HMMs 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 [135] 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 is also work on cognitive assistive systems for the elderly by Liao *et al.* [136] 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. SUMMARY

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 priori* 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. Another open area of research is object recognition, which also plays a major role in this context. During the recent years there is a more common understanding between researchers on which properties of physical objects to represent to facilitate the recognition process. Apart from a common understanding for representation, the understanding of action requires also reasoning about qualitative temporal relationships. This is why considerable research will be necessary to fully understand the problems associated with action understanding.

REFERENCES

1. M. Giese and T. Poggio, Neural mechanisms for the recognition of biological movements. *Nat. Rev.* **4**, 179–192 (2003).
2. G. Rizzolatti, L. Fogassi and V. Gallese, Parietal cortex: from sight to action, *Curr. Opin. Neurobiol.* **7**, 562–567 (1997).

3. G. Rizzolatti, L. Fogassi and V. Gallese, Neurophysiological mechanisms underlying the understanding and imitation of action, *Nat. Rev.* **2**, 661–670 (2001).
4. B. Dariush, Human motion analysis for biomechanics and biomedicine, *Machine Vis. Applic.* **14**, 202–205 (2003).
5. S. Schaal, Is imitation learning the route to humanoid robots?, *Trends Cogn. Sci.* **3**, 233–242 (1999).
6. S. Schaal, A. Ijspeert and A. Billard, Computational approaches to motor learning by imitation, *Phil. Trans. R. Soc. Lond. B* **358**, 537–547 (2003).
7. H.-H. Nagel, From image sequences towards conceptual descriptions, *Image Vis. Comput.* **6**, 59–74 (1988).
8. A. Bobick, Movement, activity, and action: the role of knowledge in the perception of motion, *Phil. Trans. R. Soc. London* **352**, 1257–1265 (1997).
9. J. Aggarwal and S. Park, Human motion: modeling and recognition of actions and interactions, in: *Proc. 2nd Int. Symp. on 3D Data Processing, Visualization and Transmission*, Thessaloniki, pp. 640–647 (2004).
10. J. Gonzàlez, J. Varona, F. Roca and J. Villanueva, *aSpaces*: action spaces for recognition and synthesis of human actions, in: *Proc. Int. Workshop on Articulated Motion and Deformable Objects*, Palma de Mallorca, pp. 189–200 (2002).
11. 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 Southern California (2002).
12. J. K. Aggarwal and Q. Cai, Human motion analysis: a review, *Comput. Vis. Image Understand.* **73**, 428–440 (1999).
13. D. M. Gavrila, The visual analysis of human movement: a survey, *Comput. Vis. Image Understand.* **73**, 82–98 (1999).
14. Y. Wu and T. S. Huang, Vision-based gesture recognition: a review, *Lecture Notes Comput. Sci.* **1739**, 103–116 (1999).
15. T. Moeslund, A. Hilton and V. Krueger, A survey of advances in vision-based human motion capture and analysis, *Comput. Vis. Image Understand.* **104**, 90–127 (2006).
16. A. Veeraranghavan, R. Chellappa and A. Roy-Chowdhury, The function space of an activity, in: *Proc. Computer Vision and Pattern Recognition*, New York, NY, pp. 17–22 (2006).
17. Y. Kuniyoshi, M. Inaba and H. Inoue, Learning by watching, extracting reusable task knowledge from visual observation of human performance, in: *IEEE Trans. Robotics Automat.* **10**, 799–822 (1994).
18. A. Billard, Imitation: a review, in: *Handbook of Brain Theory and Neural Network* (M. Arbib, Ed.), pp. 566–569. MIT Press, Cambridge, MA (2002).
19. K. Ogawara, S. Iba, H. Kimura and K. Ikeuchi, Recognition of human task by attention point analysis, in: *Proc. IEEE Int. Conf. on Intelligent Robot and Systems*, Takamatsu, pp. 2121–2126 (2000).
20. K. Ogawara, S. Iba, H. Kimura and K. Ikeuchi, Acquiring hand-action models by attention point analysis, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Seoul, pp. 465–470 (2001).
21. M. C. Lopes and J. S. Victor, Visual transformations in gesture imitation: what you see is what you do, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Taipei, pp. 2375–2381 (2003).
22. O. C. Jenkins and M. J. Mataric, Performance-derived behavior vocabularies: data-driven acqusition of skills from motion, *Int. J. Humanoid Robotics* **1**, 237–288 (2004).
23. S. Ekvall and D. Kragic, Grasp recognition for programming by demonstration tasks, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, pp. 748–753 (2005).
24. S. Calinon, F. Guenter and A. Billard, Goal-directed imitation in a humanoid robot, in: *Proc. Int. Conf. on Robotics and Automation*, Barcelona (2005).
25. G. Rizzolatti, L. Fadiga, V. Gallese and L. Fogassi, Premotor cortex and the recognition of motor actions, *Brain Res. Cogn. Brain Res.* **3**, 131–141 (1996).

26. V. Ramachandran, Mirror neurons and imitation learning as the driving force behind the gerat leap forward in human evolution, *Edge* **69** (2000).
27. L. Fadiga, L. Fogassi, V. Gallese and G. Rizzolatti, Visuomotor neurons: ambiguity of the discharge or 'motor perception'?, *Int. J. Psychophysiol.* **35**, 165–177 (2000).
28. G. Rizzolatti, L. Fogassi and V. Gallese, Motor and cognitive functions of the ventral premotor cortex, *Curr. Opin. Neurobiol.* **12**, 149–154 (2002).
29. G. Metta, G. Sandini, L. Natale, L. Craighero and L. Fadiga, Understanding mirror neurons: a bio-robotic approach, *Interact. Studies* **7**, 197–232 (2006).
30. D. Ormoneit, H. Sidenbladh, M. Black and T. Hastie, Learning and tracking cyclic human motion, in: *Proc. Workshop on Human Modeling, Analysis and Synthesis at CVPR*, Hilton Head Island, SC (2000).
31. H. Sidenbladh, M. Black and L. Sigal, Implicit probabilistic models of human motion for synthesis and tracking, in: *Proc. Eur. Conf. on Computer Vision*, Copenhagen, pp. 784–800 (2002).
32. J. Deutscher, A. Blake and I. Reid, Articulated body motion capture by annealed particle filtering, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Hilton Head Island, SC, Volume II, pp. 126–133 (2000).
33. A. Wilson and A. Bobick, Parametric hidden Markov models for gesture reocngition, *IEEE Trans. Pattern Anal. and Machine Intell.* **21**, 884–900 (1999).
34. M. C. Lopes and J. S. Victor, Visual learning by imitation with motor representations, *IEEE Trans. Syst. Man Cybernet. B* **35**, 438–449 (2005).
35. M. C. Lopes and J. S. Victor, A developmental roadmap for learning by imitation in robots, *IEEE Trans. Syst. Man Cybernet. B* **37**, 308–321 (2007).
36. N. Robertson and I. Reid, Behaviour understanding in video: a combined method, in: *Proc. Int. Conf. on Computer Vision*, Beijing, pp. 808–815 (2005).
37. K. Sato and J. Aggarwal, Tracking and recognizing two-person interactions in outdoor image sequences, in: *Proc. Workshop on Multi-Object Tracking*, Vancouver, pp. 87–94 (2001).
38. K. Ikeuchi and T. Suehiro, Towards assembly plan from observation, part I: Task recognition with polyhedral objects, *IEEE Trans. Robotics Automat.* **10**, 368–385 (1994).
39. M. Steedman, Plans, affordances, and combinatory grammar, *Linust. Philos.* **25**, 723–753 (2002).
40. M. Steedman, Temporality, in: *Handbook of Logic and Language* (J. van Benthem and A. ter Meulen, Eds), pp. 895–938. Elsevier, Amsterdam (1997).
41. P. Fitzpatrick, G. Metta, L. Natale, S. Rao and G. Sandini, Learning about objects through action—initial steps towards artificial cognition, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Taipei, pp. 3140–3145 (2003).
42. Y. Demiris and M. Johnson, Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning, *Connect. Sci. J.* **15**, 231–243 (2003).
43. C. Stauffer and W. Grimson, Learning patterns of activity using real-time tracking, *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 747–757 (2000).
44. O. Boiman and M. Irani, Detecting irregularities in images and in video, in: *Proc. Int. Conf. on Computer Vision*, Beijing, pp. 462–469 (2005).
45. A. Chowdhury and R. Chellappa, A factorization approach for activity recognition, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Washington, DC, pp. 22–31 (2004).
46. N. Vasvani, A. R. Chowdhury and R. Chellappa, Activity recognition using the dynamics of the configuration of interacting objects, in: *Proc. Computer Vision and Pattern Recognition*, Madison, WI, pp. 633–640 (2003).
47. D. Kendall, D. Barden, T. Carne and H. Le, *Shape and Shape Theory*. Wiley, New York, NY (1999).
48. C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: a factorization method, *Int. J. Comput. Vis.* **9**, 137–154 (1992).

49. T. Xiang and S. Gong, Beyond tracking: modelling action and understanding behavior, *Int. J. of Comput. Vis.* **67**, 21–51 (2006).
50. A. Efros, A. Berg, G. Mori and J. Malik, Recognizing action at a distance, in: *Proc. Int. Conf. Computer Vision*, Nice, Volume II, pp. 726–733, (2003).
51. Y. Ricquebourg and P. Bouthemy, Real-time tracking of moving persons by exploiting spatio-temporal image slices, *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 797–808 (2000).
52. J. Rittscher, A. Blake and S. Roberts, Towards the automatic analysis of complex human body motions, *Image Vis. Comput.* **20**, 905–916 (2002).
53. A. Yilmaz and M. Shah, Actions sketch: a novel action representation, in: *Proc. Computer Vision and Pattern Recognition*, San Diego, CA, pp. 984–989 (2005).
54. B. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, in: *Proc. Int. Conf. on Computer Vision*, Beijing, pp. 1395–1402 (2005).
55. L. Gorelick, M. Galun, E. Sharon, A. Brandt and R. Basri, Shape representation and recognition using the Poisson equation, in: *Proc. Computer Vision and Pattern Recognition*, Washington, DC, Volume 2, pp. 61–67 (2003).
56. H. Yu, G.-M. Sun, W.-X. Song and X. Li, Human motion recognition based on neural networks, in: *Proc. Int. Conf. on Communications, Circuits and Systems*, Hong Kong, Volume 2, pp. 982–989 (2005).
57. M. Rahman and A. Robles-Kelly, A tuned eigenspace technique for articulated motion recognition, in: *Proc. Eur. Conf. on Computer Vision*, Graz (2006).
58. H. Jiang, M. Drew and Z. Li, Successive convex matching for action detection, in: *Proc. Computer Vision and Pattern Recognition Conf.*, New York, NY, pp. 1646–1653 (2006).
59. A. Bobick and J. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Anal. Machine Intell.* **23**, 257–267 (2001).
60. G. Bradski and J. Davis, Motion segmentation and pose recognition with motion history gradients, *Machine Vis. Applic.* **13**, 174–184 (2002).
61. A. Elgammal and C. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Washington, DC, Volume 2, pp. 681–688 (2004).
62. A. Elgammal and C. Lee, Separating style and content on a nonlinear manifold, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Washington, DC, Volume I, pp. 478–485 (2004).
63. S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**, 2323–2327 (2000).
64. J. Tenenbaum, V. de Silva and J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319–2323 (2000).
65. J. Yamato, J. Ohya and K. Ishii, Recognizing human action in time-sequential images using hidden Markov model, in: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Champaign, IL, pp. 379–385 (1992).
66. A. Elgammal, V. Shet, Y. Yacoob and L. Davis, Learning dynamics for exemplar-based gesture recognition, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Madison, WI, Volume I, pp. 571–577 (2003).
67. 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, *Comput. Vis. Image Under.* **92**, 196–216 (2003).
68. Y. Shi, A. Bobick and I. Essa, Learning temporal sequence model from partially labeled data, in: *Proc. Computer Vision and Pattern Recognition*, New York, NY, pp. 1631–1638 (2006).
69. J. Davis and S. Taylor, Analysis and recognition of walking movements, in: *Proc. Int. Conf. Pattern Recognition*, Quebec, Volume I, pp. 315–318 (2002).
70. I. Haritaoglu, D. Harwood and L. Davis, W4: real-time surveillance of people and their activities, *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 809–830 (2000).

71. 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, CA, pp. 941–959 (2004).
72. A. Gritai, Y. Sheikh and M. Shah, On the use of anthropometry in the invariant analysis of human actions, in: *Proc. Int. Conf. on Pattern Recognition*, Cambridge (2004).
73. Y. Sheikh, M. Sheikh and M. Shah, Exploring the space of human action, in: *Proc. Int. Conf. on Computer Vision*, Beijing, pp. 144–149 (2005).
74. C. Fanti, L. Zelnik-Manor and P. Perona, Hybrid models for human motion recognition, in: *Proc. Computer Vision and Pattern Recognition Conf.*, San Diego, CA, pp. 1166–1173 (2005).
75. A. Yilmaz and M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: *Proc. Int. Conf. on Computer Vision*, Beijing, pp. 150–157 (2005).
76. J. Davis and H. Gao, Gender recognition from walking movements using adaptive three-mode PCA, in: *Proc. Computer Vision and Pattern Recognition Conf.*, Washington, DC, pp. 9-9 (2004).
77. V. Parameswaran and R. Chellappa, View invariance for human action recognition, *Int. J. Comput. Vis.* **66**, 83–101 (2006).
78. H. Ren and G. Xu, Human action recognition with primitive-based coupled-HMM, in: *Proc. Int. Conf. on Pattern Recognition*, Quebec, pp. 494–498 (2002).
79. H. Ren, G. Xu and S. Kee, Subject-independent natural action recognition, in: *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Seoul, pp. 523–528 (2004).
80. D. Vecchio, R. Murray and P. Perona, Decomposition of human motion into dynamics-based primitives with application to drawing tasks, *Automatica* **39**, 2085–2098 (2003).
81. C. Lu and N. Ferrier, Repetitive motion analysis: segmentation and event classification, *IEEE Trans. Pattern Anal. Machine Intell.* **26**, 258–263 (2004).
82. C. Rao, A. Yilmaz and M. Shah, View-invariant representation and recognition of actions, *J. Comput. Vis.* **50**, 203–226 (2002).
83. S. Park and J. Aggarwal, Semantic-level understanding of human actions and interactions using event hierarchy, in: *Proc. CVPR Workshop on Articulated and Non-Rigid Motion*, Washington, DC, pp. 12-12 (2004).
84. Y. Ivanov and A. Bobick, Recognition of visual activities and interactions by stochastic parsing, *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 852–872 (2000).
85. A. Stolcke, An efficient probabilistic context-free parsing algorithm that computes prefix probabilities, *Computat. Linguist.* **21**, 165–201 (1995).
86. V. Krueger and D. Grest, Using hidden Markov models for recognizing action primitives in complex actions, in: *Proc. Scandinavian Conf. on Image Analysis*, Aalborg, pp. 203–213 (2006).
87. M. Yamamoto, H. Mitomi, F. Fujiwara and T. Sato, Bayesian classification of task-oriented actions based on stochastic context-free grammar, in: *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Southampton (2006).
88. F. Lv and R. Nevatia, Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost, in: *Proc. Eur. Conf. on Computer Vision*, Graz, pp. 359–372 (2006).
89. C. Breazeal and B. Scassellati, Robots that imitate humans, *Trends Cognitive Sci.* **6**, 481–487 (2002).
90. R. Dillmann, Teaching and learning of robot tasks via observation of human performance, *Robotics Autonomous Syst.* **47**, 109–116 (2004).
91. A. Stoytchev, Behavior-grounded representation of tool affordances, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, pp. 3060–3065 (2005).
92. D. Newton, D. Engquist and J. Boos, The objective basis of behavior unit, *J. Personal. Social Phsychol.* **35**, 847–862 (1977).

93. M. A. Arbib, Perceptual structures and distributed motor control, in: *Handbook of Physiology, Section 2: The Nervous System (Vol. II, Motor Control, Part 1)* (V. B. Brooks, Ed.), pp. 1449–1480. American Physiological Society, Washington, DC (1981).
94. M. J. Mataric, M. Williamson, J. Demiris and A. Mohan, Behavior-based primitives for articulated control, in: *Proc. 5th Int. Conf. on the Simulation of Adaptive Behavior*, Cambridge, MA, pp. 165–170 (1998).
95. S. Schaal, Dynamic movement primitives—a framework for motor control in humans and humanoid robotics, in: *Proc. Int. Symp. on Adaptive Motion of Animals and Machines*, pp. 12–20 (2003).
96. R. W. Paine and J. Tani, Motor primitive and sequence self-organization in a hierarchical recurrent neural network, *Neural Networks* **17**, 1291–1309 (2004).
97. A. Billard, Y. Epars, S. Calinon, S. Schaal and G. Cheng, Discovering optimal imitation strategies, *Robotics Autonomous Syst.* **47**, 69–77 (2004).
98. T. Inamura, I. Toshima, H. Tanie and Y. Nakamura, Embodied symbol emergence based on mimesis theory, *Int. J. Robotics Res.* **23**, 363–377 (2004).
99. F. A. Mussa-Ivaldi and E. Bizzi, Motor learning through the combination of primitives, *Phil. Trans. R. Soc. Lond. B* **355**, 1755–1769 (2000).
100. A. IJspeert, J. Nakanishi and S. Schaal, Movement imitation withnonlinear dynamical systems in humanoid robots, in: *Proc. Int. Conf. on Robotics and Automation*, Washington, DC, pp. 1398–1403 (2002).
101. D. C. Bentivegna, C. G. Atkeson, A. Ude and G. Cheng, Learning to act from observation and practice, *Int. J. Humanoid Robotics* **1**, 585–611 (2004).
102. C. Nehaniv and K. Dautenhahn, Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications, in: *Learning Robots: An Interdisciplinary Approach* (J. Demiris and A. Birk, Eds), pp. 136–161. World Scientific Press, Singapore (1999).
103. S. B. Kang and K. Ikeuchi, Toward automatic robotic instruction from perception—temporal segmentation of tasks from human hand motion, *IEEE Trans. Robotics Automat.* **11**, 670–681 (1995).
104. S. B. Kang and K. Ikeuchi, Toward automatic robotic instruction from perception—mapping human grasps to manipulator grasps, *IEEE Trans. Robotics Automat.* **13**, 81–95 (1997).
105. A. Ude, Robust estimation of human body kinematics from video, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Kyongju, pp. 1489–1494 (1999).
106. 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: *Proc. AAAI and CMU Workshop on Interactive Robotics and Entertainment*, Pittsburgh, PA, pp. 35–42 (2000).
107. A. Ude, C. G. Atkeson and M. Riley, Programming full-body movements for humanoid robots by observation, *Robotics Autonomous Syst.* **47**, 93–108 (2004).
108. M. Ruchanurucks, S. Nakaoka, S. Kudo and K. Ikeuchi, Humanoid robot motion generation with sequential physical constraints, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Orlando, FL, pp. 2649–2654 (2006).
109. Y. Kuniyoshi, Y. Yorozu, M. Inaba and H. Inoue, From visuo-motor self learning to early imitation—a neural architecture for humanoid learning, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Taipei, pp. 3132–3139 (2003).
110. A. N. Meltzoff and M. K. Moore, Imitation of facial and manual gestures by human neonates, *Science* **198**, 75–78 (1977).
111. 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).
112. C. G. Atkeson and S. Schaal, Robot learning from demonstration, in: *Proc. 14th Int. Conference on Machine Learning*, pp. 12–20. Morgan Kaufmann, San Mateo, CA (1997).

113. J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal and M. Kawato, Learning from demonstration and adaptation of biped locomotion, *Robotics and Autonomous Syst.* **47**, 79–91 (2004).
114. 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 (2006).
115. C. Sminchiescu, A. Kanaujia, Z. Li and D. Metaxas, Conditional models for contextual human motion recognition, in: *Proc. Int. Conf. on Computer Vision* (ICCV'05), Beijng, pp. 1808–1815 (2005).
116. K. Ogawara, J. Takamatsu, K. Kimura and K. Ikeuchi, Generation of a task model by intergrating multiple observations of human demonstrations, in: *Proc. IEEE Int. Conf. on Robotics and Automation* (ICRA'02), Washington, DC, pp. 1545–1550 (2002).
117. S. Ekvall and D. Kragic, Learning task models from multiple human demonstrations, in: *Proc. 15th IEEE Int. Symp. on Robot and Human Interactive Communication*, Hatfield, pp. 358–363 (2006).
118. M. Johnson and Y. Demiris, Hierarchies of coupled inverse and forward models for abstraction in robot action planning, recognition and imitation, in: *Proc. AISB Symp. on Imitation in Animals and Artifacts*, Newcastle-upon-Tyne, pp. 69–76 (2005).
119. C. Schmidt, N. Sridharan and J. Goodson, The plan recognition problem: an intersection of psychology and artificial intelligence, *Artificial Intell.* **11**, 45–83 (1978).
120. R. Wilensky, *Planning and Understanding*. Addison-Wesley, Reading, MA (1983).
121. P. R. Cohen, C. R. Perrault and J. F. Allen, Beyond question answering, in: *Strategies for Natural Language Processing* (W. Lehnert and M. Ringle, Eds), pp. 245–274. Lawrence Erlbaum, Hillsdale, NJ (1981).
122. H. Kautz and J. F. Allen, Generalized plan recognition, in: *Proc. Conf. of the American Association of Artificial Intelligence*, Pittsburgh, PA, pp. 32–38 (1986).
123. J. McCarthy, Circumscription—a form of non-monotonic reasoning, *Artificial Intell.* **13**, 27–39, 171–172 (1980).
124. H. Kautz, A formal theory of plan recognition and its implementation, *PhD Thesis*, University of Rochester (1991).
125. E. Charniak and R. P. Goldman, A Bayesian model of plan recognition, *Artificial Intell.* **64**, 53–79 (1993).
126. E. Charniak and D. McDermott, *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA (1987).
127. M. Vilain, Deduction as parsing, in: *Proc. Conf. of the American Association of Artifical Intelligence*, Anaheim, CA, pp. 464–470 (1991).
128. C. L. Sidner, Plan parsing for intended response recognition in discourse, *Computat. Intell.* **1**, 1–10 (1985).
129. D. Pynadath and M. Wellman, Probabilistic state-dependent grammars for plan recognition, in: *Proc. Int. Conf. on Uncertainty in AI*, pp. 507–514 (2000).
130. R. P. Goldman, C. W. Geib and C. A. Miller, A new model of plan recognition, in: *Proc. Conf. on Uncertainty in Artificial Intelligence*, pp. 245–254 (1999).
131. C. W. Geib and R. P. Goldman, Recognizing plan/goal abandonment, in: *Proc, Int. Joint Conf. on Artificial Intelligence* (2003).
132. C. Geib, Plan recognition, in: *Adversarial Reasoning* (A. Kott and W. McEneaney, Eds), pp. 77–95. Chapman and Hall/CRC, London (2007).
133. J. G. E. Barton, On the complexity of ID/LP parsing, *Computat. Linguist.* **11**, 205–218 (1985).
134. D. Avrahami-Zilberbrand and G. A. Kaminka, Fast and complete symbolic plan recognition, in: *Proc. Int. Joint Conf. on Artificial Intelligence*, Edinburgh, pp. 173–180 (2005).
135. H. H. Bui, S. Venkatesh and G. West, Policy recognition in the abstract hidden Markov model, *J. AI Res.* **17**, 451–499 (2002).

136. L. Liao, D. Fox and H. A. Kautz, Location-based activity recognition using relational Markov networks, in: *Proc. Int. Joint Conf. on Artificial Intelligence*, Edinburgh, pp. 773–778 (2005).
137. D. Herzog, V. Krueger and D. Grest, Exemplar-based parametric hidden Markov models for recognition and synthesis of movements, in: *Proc. Int. Workshop on Vision, Modeling, and Visualization (VMV07)*, Saarbrücken, Nov. 7–9, in press (2007).

ABOUT THE AUTHORS

Volker Krüger studied Computer Science and Mathematics at the University of Kiel, where he received his MSc, in 1996 and his PhD degree, in 2000. He joint the Center for Automation Research at the University of Maryland as a Post-doc where he was the Lead Scientist for the HumanID project. Since 2002, he is an an Associate Professor at Aalborg University and Head of the Computer Vision and Machine Intelligence Lab. His research interests include computer vision, surveillance and machine learning.

Danica Kragic received the MSc degree in Mechanical Engineering from the Technical University of Rijeka, Croatia and PhD degree in Computer Science from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1995 and 2001, respectively. She is currently an Assistant Professor in Computer Science at KTH and Chairs the IEEE RAS Committee on Computer and Robot Vision. She received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. Her research interests include computer vision, service robotics and human–robot interaction.

Ales Ude studied Applied Mathematics at the University of Ljubljana, Slovenia, and Computer Science at the University of Karlsruhe, Germany, where he received a doctoral degree. He was an STA fellow in the Kawato Dynamic Brain Project, ERATO, JST. Later he was associated with the ICORP Computational Brain Project, Japan Science and Technology Agency and ATR Computational Neuroscience Laboratories, Kyoto, Japan. He is also associated with the Jozef Stefan Institute, Ljubljana, Slovenia. His research focuses on humanoid robot vision, visual perception and learning of human activity and humanoid cognition.

Christopher Geib is a Research Fellow in the School of Informatics at the University of Edinburgh, UK. His principle research interests are in Artificial Intelligence methods for plan and intent recognition, and more generally grammar-based reasoning about actions under conditions of uncertainty and the interface between continuous control systems and logic-based reasoning systems. He earned his PhD from the University of Pennsylvania, in 1995, and held a Post-doctoral fellowship at the University of British Columbia until 1997. In 1997, he became a Senior Research scientist at Honeywell Labs, where he remained until he joined the faculty at Edinburgh in 2006.