Handbook of Robotics

Chapter 64: Perceptual Robotics: Example-based representations of shapes and movements

Heinrich H. Bülthoff, Christian Wallraven
Max Planck Institute for Biological Cybernetics
Dept. of Cognitive and Computational Psychophysics
Spemannstr. 38, 72076 Tübingen
GERMANY

Martin A. Giese
School of Psychology, Univ. of Wales, Bangor
Bangor, Gwynedd LL57 2AS, UK &
ARL, Hertie Institute for Clinical Brain Research
Frondsbergstr. 23, 72076 Tübingen, GERMANY

June 15, 2007

64.1. Introduction

The technical realization of perceptual functions is a central problem for many applications in robotics. Robots require perception to navigate in space and to localize and recognize goal objects, e.g. for manipulation (e.g. Chapters 15, 28, 35, 41). Social interactive robots must be able to interpret gestures, actions, and even emotions (cf. Chapter 59) in order to interact naturally with their users. One important approach for the programming of complex perceptual and behavioral functions such as, for example, needed for humanoid robots is imitation learning (Chapter 60, 63). Imitation learning requires the robot to 'perceive' complex actions that are executed by the user and to subsequently map it into an efficient representation that is suitable for the synthesis of the corresponding motor behavior on the available platform. This chapter focuses on important principles of the representation of complex shapes and movements, which can be derived from biological perception systems, and more specifically the basic functionality of the primate visual cortex. Such principles have interesting implications for the design of technical systems in robotics and computer vision for the *recognition of objects, shapes and faces*, and for the recognition and synthesis of *complex movements and actions*. The limited space of the chapter forced us to focus mainly on visual perception and related technical applications. In the context of robotics many other aspects of perception are important, for example haptic perception (cf. Chapter 31), auditory perception, sensory cue fusion (Chapter 25), or the interaction between the visual recognition of objects and actions and motor programs, e.g. during grasping (specifically treated in Chapter 63).

In the following, we will first formulate several biological principles that are relevant for form and motion representations, specifically in the visual system. We will then, on the one hand, describe technical systems that implement these principle using neural mechanisms that are inspired by the basic architecture of the brain. On the other hand, we will discuss also implementations that are inspired by biological principles on a more abstract level, and which exploit instead of neural networks more efficient technical algorithms for the realization of biologically relevant functions. Many of these systems are derived in the field of computer vision and are based on the advantages and limitations of modern digital computers in order to more efficiently realize biological principles of information processing.

Our approach to establish relationships between biological perception and robotics systems at different levels reflects David Marr's classical distinction of multiple levels of description, originally developed for the analysis of

vision systems (Marr, 1982): Robotics systems can be inspired by biological system at the level of *implementation,* i.e. one can try to build robots containing neural mechanisms that imitate the function of neurons in central nervous systems of biological organisms. This type of analogy between technical and biological systems coincides with the definition of 'Neurorobotics' given in Chapter 63. A transfer of principles from biological perception systems to robots might also be accomplished at the more abstract levels of *computational problems* and *algorithms*. The computational level is defined by the abstract theoretical formulation of computational problems that have to be solved by perception systems. Examples are the identification or classification of goal object, or the recognition of human gestures. Marr's level of algorithms specifies the computational methods for the solution of such problems, independent of the underlying specific hardware or architecture. For example, an object might be represented by modeling its full 3D structure, e.g. using a parametric 3D shape model, or it might be represented in terms of two-dimensional example views. Example views, however, might be represented using neural networks, establishing an analogy with the human brain at the level of implementation, or using more efficient computational methods, e.g. as support vectors of a classifier that has been trained with appropriate images of the object and distractor patterns. In both cases, the robot system realizes mechanisms that are derived form perception in biological systems.

Marr's distinction of levels is only one way to introduce description levels for complex systems. Other approaches, particularly relevant for robotics, are for example the subsumption architecture (Brooks, 1986) and behavior-based approaches (Chapter 39) that decompose robotics system into a system of simpler behavioral modules. Another example are dynamical systems approaches to robotics (e.g. Kelso, 1995; Schöner et al. 1997; Tani & Ito, 2003) that are based on the biologically motivated idea that behaviors can be mapped onto stable states of (nonlinear) dynamical systems or recurrent neural networks. Individual behaviors result by self-organization over the whole system as collectively stable modes, which can be described and analyzed by the introduction of appropriate collective variables. Interestingly, such robotics-inspired approaches have been quite successful in modeling human navigation behavior (Warren, 2006).

In the following we will apply the term *Perceptual Robotics* to signify the design of robots based on principles that are derived from human perception on all three levels in the sense of Marr. This includes a realization in terms of specific neural circuits as well as the transfer of more abstract biologically-inspired strategies for the solution of relevant computational problems. A direct interaction between robotics and perception research can be very fruitful for both disciplines. On the one hand, our current knowledge about the human perception and the underlying computational principles might help to build more efficient robotics architectures that inherit properties from biological perception, e.g. very efficient and robust processing or complex dynamic flexibility. Such architectures will be a necessary pre-requisite for the creation of truly intelligent, cognitive robots (Chapters 9, 39, 59). On the other hand, perception science often uses robots as testbed for gaining a deeper understanding of computational processes, an in particular for testing the computational power of specific computational solutions under 'real-world conditions'. How can a child, for example, learn how to handle new objects, and what allows us to learn the visual categorization of thousands of objects from just a few examples? 'Perceptual robot platforms', equipped with a variety of sensory inputs and operating in different types of artificially structured or real-world environments provide very helpful tools for the study of such questions.

64.2. Example-based object representations

Object recognition is a fundamental visual function that is critical for many applications in robotics. Manipulation and grasping (Chapters 26 and 28) require exact knowledge about the shape of the goal object that often is derived from visual sensors. Also the imitation of goal directed movements (Chapter 60) requires knowledge about target objects. Finally, social and collective robots require robust recognition of other agents and objects which are taking part in the present action (Chapters 41 and 59). The importance of object and shape recognition for many other applied robot systems, like construction and assembly robots or smart cars (Chapter 52) is immediately evident.

.

64.2.1 Perceptual and computational basis of object representations

The question of how humans learn, represent and recognize objects under a wide variety of viewing conditions presents a great challenge to both neurophysiology and cognitive research. Frameworks for explaining the amazing robustness of human recognition processes and how humans represent objects can be broadly classified into two approaches: in the *model-based representation*, an image on the retina is analyzed to yield three-dimensional parts of an object based on geometric primitives (cf. also Chapter 23). These primitives are then matched to an internal, three-dimensional model of the object (see Figure 1, bottom). *Exemplar-based representation* approaches assume that the internal storage consists of, typically two-dimensional, snapshot-like representations of objects, which are directly compared to the visual input via simple image transformations. In the following, we will briefly describe the basic properties of these two approaches as well as perceptual evidence for their plausibility in explaining human recognition performance.

Structural description models

The basic idea of structural description models is that object recognition or categorisation is based on a structural representation, which is defined as a configuration of elementary object parts that are regarded as *shape primitives* (e.g., Marr & Nishihara, 1978). Structural description models aim at supplying abstract and propositional descriptions of objects, while at the same time disregarding irrelevant spatial information. Therefore, structural description models typically predict that recognition performance is invariant regarding spatial transformations. Biederman's recognition-by-components (RBC) or geon structural description (GSD) model can be regarded as the best developed example of the structural description model type (Biederman, 1987). According to this model objects are represented as configurations of elementary three-dimensional primitive parts, called geons. These geons are derived from nonaccidental properties (NAPs) in the image, i.e. from properties which unlikely arise by chance, and are more or less invariant over a wide range of views. For example, the properties straight vs. curved, symmetrical vs. asymmetrical, parallel vs. nonparallel are regarded as nonaccidental properties (nonaccidental properties were originally proposed within an image-based approach by Lowe, 1985). According to the model, geons and their spatial configuration are combined into a structural representation, called geon structural description. The spatial relations between parts are described in a categorical way, using relations like *above*, *below*, etc. Like other structural description models, Biederman's model predicts invariance in relation to position and size and also in relation to orientation in depth, as long as no parts are occluded.

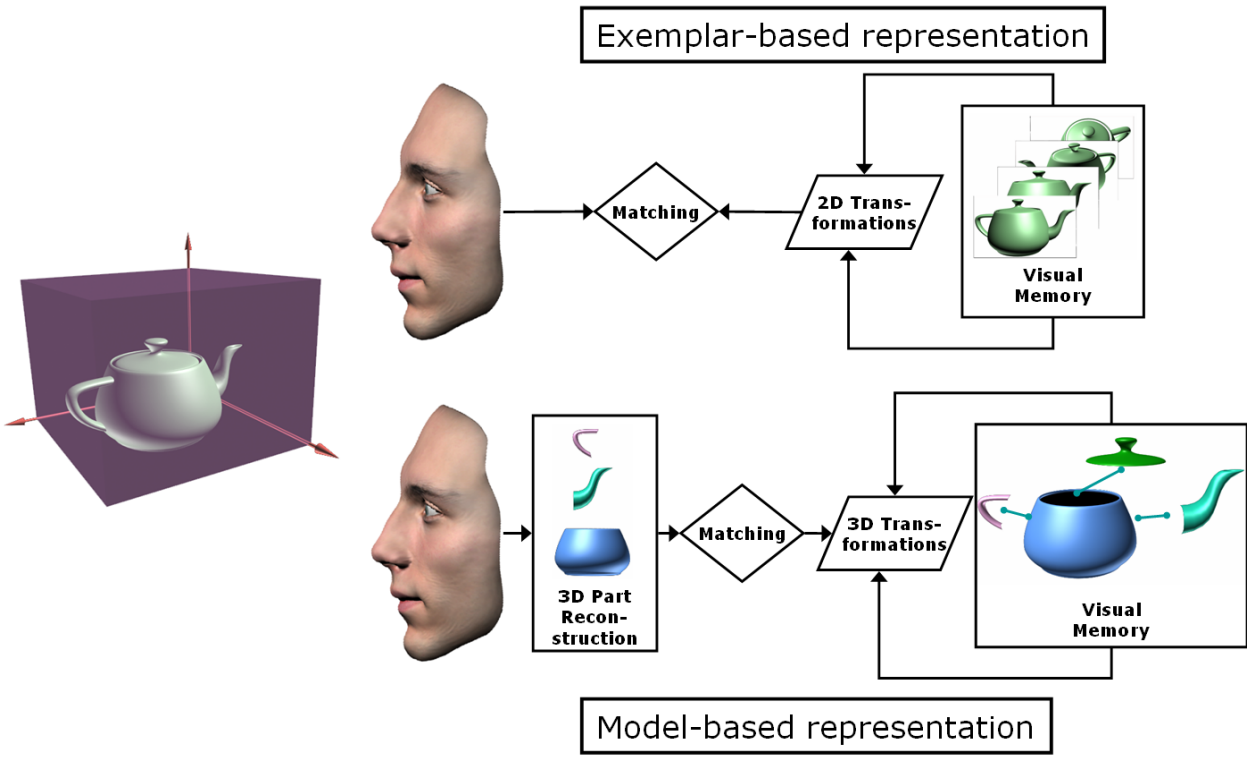

Figure 1. Schematic drawing comparing exemplar-based with model-based representations. Object perception based on model-based representations assumes that the brain extracts 3D-parts from the visual image, which are then matched to an internally stored 3D-model of the teapot. Contrasting with this approach, object perception based on exemplar-based representations is accomplished by directly comparing stored templates or example images with the current picture of the teapot.

The question has to be raised whether objects can be decomposed into geons at all. It was argued that Biederman's RBC cannot be applied to a whole range of biological stimuli (Ullman, 1996), or that biological shapes in general cannot be adequately described by structural description models (Kurbat, 1994). This problem extends also to artefact categories like *shoe*, *hat* or *backpack*, which seem to exceed the scope of the geon model. Therefore it has to be doubted that object parts are necessarily represented as geons, or as similar geometrical primitives (for further problems of RBC see Edelman, 1999; Tarr & Bülthoff, 1998). However, this does not mean that category representations do not have a part structure: in fact it is not the notion of part structure in object representations by itself which is problematic, but the use of parts and relations as a basis to derive invariant recognition performance (see Graf & Schneider, 2001).

Exemplar-based models

Over the last two decades, an increasing number of studies has demonstrated that recognition is not view-independent. Orientation dependent recognition effects were found for novel objects (e.g., Bülthoff & Edelman, 1992; Tarr & Pinker, 1989), and also for common, familiar objects (e.g., Hayward & Tarr, 1997; Palmer, Rosch & Chase, 1981). Orientation-dependent recognition performance has been shown not to be limited to individual

objects, such as faces (e.g., Hill, Schyns & Akamatsu, 1997; Wallraven et al., 2002), or to objects on the subordinate level of categorisation (e.g., Edelman & Bülthoff, 1992; Tarr, 1995), but also was demonstrated for basic level recognition (Lawson & Humphreys, 1998; Palmer, Rosch & Chase, 1981).

Moreover, recognition performance is not only influenced by the orientation, but also by the *size* of the stimulus. Results are quite similar: Reaction times (RTs) and error rates depend on the extent of transformation that is necessary to align memory and stimulus representation. RTs increase in a monotonic way with increasing change of (perceived) size (for a review see Ashbridge & Perrett, 1998). Several studies even show a systematic relationship between the amount of *translation* and recognition performance: Increasing displacement between two sequentially presented stimuli led to a deterioration of performance, both for novel objects (Dill & Edelman, 2001) and familiar objects (Cave et al., 1994). Overall, view-independent models are difficult to reconcile with these findings which indicate that recognition performance depends systematically on different spatial transformations.

In the following, we will briefly review three types of exemplar-based models, which – by virtue of different computational mechanisms and processes (including alignment, interpolation, and pooling/thresholding) – explain the transformation-dependent performance that was found in the psychophysical experiments.

In the class of *alignment models*, Ullman's (1989, 1996) 3D alignment model and Lowe's (1985) SCERPO model are probably the best known examples. Both models work by storing 3D models of objects, which are aligned to images by perspective projection of corresponding features (edges or feature points on the object). As an alternative to Ullman's (1989) model that relies on 3D object representations, Ullman and Basri (1991) suggested an alignment model on the basis of 2D views. In this model, an internal object model is constructed by a linear combination of a small number of stored 2D exemplar images. Thus, the alignment is not achieved by a spatial compensation process, but by linear combination of images. The intuition behind the linear combination approach can be explained in simple terms: Suppose that two views of the same three-dimensional object are stored, taken from somewhat different viewing directions. An intermediate view can then be described as a weighted sum of the views that are already stored. In this case, the representation is based on the two-dimensional positions of corresponding features in each view. Making the set of views closer results in an object representation that is equivalent to storing a 3D model.

In the *interpolation model*, recognition is achieved by localization in a multidimensional representational space, which is spanned by stored views (Poggio & Edelman, 1990). The interpolation model is based on the theory of approximation of multivariate functions and can be implemented with radial basis functions (RBFs). In this scheme, the whole viewing space of an object is approximated by the learned exemplar views through a limited number of series of so-called Radial Basis Functions (such as Gaussian functions) each of which becomes activated within a limited region of the high-dimensional feature space. Object recognition then means to examine whether a new point corresponding to the actual stimulus can be approximated by the existing tuned set of basis functions. Thus, recognition does not occur by transformation or reconstruction of an internal image, but rather by interpolation or approximation of exemplars in a high-dimensional representational space.

At the end of the 90s – and as an extension to the interpolation models – recognition models based on *pooling and thresholding* were developed (Perrett & Oram, 1998; Riesenhuber & Poggio, 1999; Rolls & Milward, 2000;

Wallis & Bülthoff, 1999). Recognition is explained on the basis of the behaviour of cells that are selectively tuned to specific image features (fragments or whole shapes) in a view-dependent (and size-dependent) way. A hierarchical pooling of the outputs of view-specific cells provides generalization over viewing conditions (Perrett & Oram, 1998). A similar proposal was made by Riesenhuber and Poggio (1999). The threshold model (Perrett, Oram & Ashbridge, 1998) also accounts for the systematic relation between recognition latencies and the amount of rotation (and size-scaling): The speed of object recognition depends on the rate of accumulation of activity from neurons selective for the object, evoked by a particular viewing circumstance. For a familiar object, more tuned cells will be activated in the views most frequently presented, so that a given level of evidence (threshold) can be achieved fast. When the object is seen in an unusual view, fewer cells will respond, and activity among the population of cells selective for the object's appearance will accumulate more slowly. Consequently, these threshold models explain orientation-dependency without the need to postulate transformation or interpolation processes.

Several computational models have been proposed that aim at modelling and explaining the dependence of human recognition performance on spatial transformations. All of these models rely on storing exemplars – in the simplest form just 2D views of objects – and matching the retinal image to these stored examples by different computational methods. The later models of recognition take their inspiration from recent findings from physiological studies concerning the functional building blocks of human vision in the brain. In the following, we will therefore briefly review the neural processing of visual information in the brain that underlies our ability to recognize objects.

64.2.2 Neural representations in object recognition

Functionally, it has been shown that the flow of visual information in the brain can be divided into two major pathways: the *dorsal pathway* is believed to process motion and motor- or action-related visual information, whereas the *ventral pathway* usually is associated with the task of object recognition. The structure of the ventral pathway is hierarchically organized and consists of a series of interconnected stages that start from the retina, passing through the Lateral Geniculate Nucleus (LGN) to the primary visual cortex (V1) and extrastriate visual areas V2, V4 and IT. The *inferotemporal cortex* (IT) provides input to the *prefrontal cortex* (PFC), which is believed to play an important role in identification and categorization of visual stimuli. Recordings in the *parietal cortex* (e.g. Sakata 2003) suggest, in addition, that specifically for grasping and object manipulation also dorsal regions might be centrally involved in the recognition of manipulable objects and their affordances (see Chapter 63 for a more detailed discussion).

The seminal work of Hubel and Wiesel (1962) in the cat (and later also in the macaque) visual cortex first established the idea of a hierarchical organization of visual processing. They found so-called simple cells in the early visual cortex (area V1) that responded best to bar-like stimuli at a particular orientation and position in the visual field. The response pattern of these cells could be modeled as a receptive field using Gabor-type functions. Later in the processing stream they found so-called complex cells which responded best to bar-like stimuli at a particular orientation nearly everywhere in the visual field – cells, which had become partially position invariant. This general idea of increasing invariance to stimulus properties with later stages of the processing stream has been

verified in further physiological studies. In general, it has been found that the receptive field of the neurons increases and that the complexity of the stimulus it responds to increases. One of the key studies about the functional role of IT regions has investigated the responses of neurons to real-world objects in anesthetized monkeys (Wang et al., 1998, see also Tanaka 1996 for a review). Although some neurons were found which responded maximally to simple bar-like stimuli, the majority of neurons in *posterior inferotemporal cortex* (PIT) preferred complex objects such as star-shapes or circles with protruding elements. Interestingly, neurons were highly sensitive to minuscule changes to these objects such as the relative orientation or thickness of the elements. On the other hand, neurons were quite insensitive to stimulus variations such as size, contrast or retinal location. These findings were taken as evidence that one of the strategies for representing objects might be to use a number of moderately complex visual elements, whose pattern of co-activation encodes the visual appearance of the stimulus. In addition, Wang et al. (1998) found neurons in *anterior inferotemporal cortex* (AIT), which responded maximally to images of whole objects such as faces or cars, indicating that already in IT object specific encodings might be present. Several other studies have also found neurons in this area which are tuned to faces, parts of faces, as well as body parts (see Grill-Spector & Malach, 2004 for a review).

In another set of experiments, Logothetis et al. (1994) found AIT neurons, which showed a strong view-based behavior for the same stimuli that were used in the study of Bülthoff & Edelman (1992), whereas they were invariant to size and location of the stimulus. Their findings provide strong evidence that a neural implementation of view-based object encoding is possible and indeed seems to be used for recognition. In addition to view-selectivity and size invariance, the investigated cells were also found to be maximally selective for the holistic stimulus rather than its constituent parts. This finding indicates that these cells might be encoding the pooled co-activation pattern of earlier PIT cells and thus form view-tuned units of recognition. It is important to stress in this context that an abstraction such as "grandmother" neurons, which specifically encode only one stimulus, does not seem plausible. Rather, the majority of neural responses in this and other experiments showed selectivity for a number of stimuli. A plausible explanation for this finding is that objects are encoded not by a single neuron but by a population code encompassing a number of neurons, which greatly increases the robustness of the representation (see also Wallis & Bülthoff, 1999).

The findings from this area of research can be summarized in a simple functional architecture: going from early stages to later stages of visual processing in a feed-forward fashion, feature complexity increases from simple edge detectors towards view-tuned, complex object cells and invariance to changes in the stimulus increases. This functional architecture is reflected not only in the object recognition framework discussed previously, but it also provides the motivation for computational vision systems that have been developed over the last few decades which will be discussed in the following.

64.2.3 Object recognition: lessons from computer vision

Computer vision started out as a subfield of artificial intelligence in the 1960s. Early work on scene understanding by Roberts (Roberts, 1965) showed how computers could “parse” worlds consisting of simple, geometric objects such as cubes, pyramids, etc. The main thrust of computer vision systems in the following decades consisted of

building algorithms for reconstructing a three-dimensional world from images – this development was further stimulated by Marr's very influential theory of vision as 3D reconstruction (Marr, 1982). This theory was built on extracting geometric primitives from images that could be mathematically described as generalized cylinders. Although the mathematical rigor of such approaches was very appealing, computational implementations turned out to have strong limitations. Extracting robust features is a necessary prerequisite for building a 3D reconstruction of the image, and finding these features proved to be hard under real-world conditions due to the enormous amount of variation in the image caused by changes in lighting, depth rotations, noise, occlusion, etc.

Parallel to the paradigm shift in human psychophysics and physiology, exemplar-based computational systems began to emerge, which for the first time showed good recognition performance under a larger range of viewing conditions. These recognition systems were based on - sometimes surprisingly simple - histograms of pixel values (Swain and Ballard, 1991), local feature detectors (Schmid and Mohr, 1997, Lowe, 2004) or on a straight-forward pixel representation of images using principle components analysis (Kirby and Sirovich, 1990). All of these recognition systems relied on a database of labeled example images, an algorithm for extracting features from these images and a suitable classification method for comparing sets of image features.

Returning to the discussion of modeling human vision, in the following we provide an exemplary review of three relatively recent recognition systems that are based on a functionally plausible, exemplar-based architecture: these are SpikeNET (Delorme & Thorpe, 2003), LeNet (LeCun et al., 2004), and a framework by Serre et al. (2005). The first system is motivated by the finding that humans are amazingly fast at categorizing images as containing an animal or a face (Thorpe et al., 1996). Typical response times for this task are so small (on the order of 100ms) that the visual signal has only time for one feed-forward pass through the visual areas of the brain (see Fig. 2) – any recurrent feedback processing would necessarily delay the decision and therefore result in longer response times. Based on this finding, a neural network architecture was designed (Delorme and Thorpe, 2003) that exploits the timing of neuronal responses (spikes) to encode visual signals using a "who fires first"- strategy. This is different from traditional neural networks in that the timing is used rather than the firing strength. An object in this system will therefore be represented by an ensemble of neurons that represents a pattern of spike responses from earlier low-level, feature extraction neurons. In their implementation these low-level neurons consist of standard Gabor-type receptive fields that are similar to the receptive fields found in the cat's visual cortex (Hubel and Wiesel, 1967). This spike time encoding allows for very fast processing of visual stimuli and has been shown to provide robust recognition results. The network architecture LeNet (LeCun et al., 2004) consists of a neural network that uses a hierarchy of layers of trainable convolutions and spatial subsampling, as well as non-linear filtering to extract features of increasingly large receptive fields, increasing complexity, and increasing robustness. Using extensive, supervised training of the full hierarchy, such a network provides a very efficient, sparse set of features for many visual recognition tasks. Finally, the network architecture by Serre et al. (2005) uses a very similar hierarchical structure of layers in which feature complexity and invariance are successively increased by linear and nonlinear pooling - its lower-level feature detectors, however, are trained in an unsupervised fashion on a large database of natural images, yielding a large set of detectors that are optimally tuned to natural image statistics. Again, performance of this model in recognition tasks has been shown to be very good – in addition, comparisons with

physiological and psychophysical experiments have shown that this framework is also capable of modeling human results from these experiments. All of these systems have been demonstrated to work on real-world vision problems (such as recognition of objects in cluttered environments) and in the case of the system by Serre et al. (2005) have been shown to even outperform other state-of-the-art computer vision algorithms. This would make them good candidates for building the visual learning and recognition system for a perceptual robot.

64.2.4 Object learning and recognition for perceptual robotics

In general, it can be said that the success of perceptually inspired recognition systems can be seen as a strong indicator for the feasibility of a data-driven, exemplar-based approach to recognition. There are three issues, however, which so far have not been addressed in any of these vision systems and which will be important both for achieving human performance in generic recognition tasks in a perceptual robotics application – as well as for a full understanding of the processes in human object recognition.

First of all, all of the above-mentioned systems are feed-forward – virtually no feedback, recurrent processing is implemented in their architecture, which makes them in a sense very similar to the simpler frog- or bee-like neural systems discussed in Chapter 63. Although there is evidence that humans solve some recognition tasks using very little feedback (see, for example, Thorpe et al., 1996), it nevertheless is a crucial component of visual processing driving, for example, attentional focus, context awareness, as well as memory and reasoning processes – basically everything that makes up visual "intelligence". Some visual attention models that are relevant for robotics systems are reviewed in Chapter 63.

Secondly, a severe limitation of most of today's artificial recognition systems is that they solely focus on the static domain of object recognition. Visual input on the retina, however, consists of dynamic changes due to object- and self-motion, non-rigid deformations of objects, articulated object motion as well as scene changes such as variations in lighting, occluding and re- and disappearing objects - where at any given point in time several of these changes can be interacting. Several psychophysical experiments, indeed suggest an important role for dynamic information, both in learning and recognition of objects (Wallis & Bülthoff, 2001; Stone, 1998, 1999; Voung et al., 2004). These results ask for an extension of current object recognition frameworks with a temporal component in order to arrive at truly spatio-temporal object representations. Combining methods from computer vision, psychophysics, and machine learning, Wallraven et al. (2001, 2006) have developed a framework that fulfills this requirement and learns spatio-temporal, exemplar-based object representations from image sequences (see Figure 3). More specifically, spatiotemporal characteristics of the visual input are integrated into a connected view-graph representation based on tracked local features. In order to provide robust classification performance, machine learning techniques are used to design efficient methods for combining Support Vector classification schemes with these local feature representations (Wallraven, 2003). In several studies it was shown that the framework achieved excellent recognition results on both highly controlled databases as well as on real-world data. The integration of spatio-temporal information provides characteristic information about dynamic visual input via the connection of views and the two-dimensional image motion of discriminative features. In addition to delivering good recognition

performance, the framework was also able to model results from psychophysical experiments on face and object recognition.

A third issue that – in our view – will be essential for designing and implementing efficient perceptual robots consists of the multi-sensory nature of our perceptual system (see also the discussion of embodied robots in Chapter 39). As an example, there is a close coupling between the human visual and haptic system – touch can provide a wealth of complementary information about an object when it is manipulated, such as its texture, its shape, its position in space relative to our body, etc. In a series of psychophysical experiments (see Newell et al. (2001)), participants had to learn views of four simple, 3D objects made of stacked toy-bricks either in the haptic modality (when they were blind-folded) or in the visual modality (without being able to touch them). Subsequently they were tested both within the same modality as well as across modalities. Recognition results showed that cross-modal recognition is possible well above chance. Not surprisingly, recognition of rotated objects in the within-modality condition was severely affected by rotation in both modalities. This shows that not only visual recognition is highly view-dependent but also that haptic recognition performance is directly affected by different viewing parameters. The results from this experiment thus support the view that haptic recognition is also mediated by exemplar-based processes.

Taken together with the keyframe framework outlined above, this cross-modal transfer might be an important reason for the excellent visual performance of human object recognition - after all, it is known that infants learn extensively by grasping and touching objects, which thus could provide a "database" of object representations for visual recognition. Using this basic perceptual principle as a motivation, Wallraven and Bülthoff (2006) have applied an extension of the keyframe framework in an on-line robotics scenario for efficient learning and recognition of multi-sensory object representations More specifically, a framework was developed to integrate both *proprioceptive* information originating from haptic sensors in the robot's hands and *visual* information coming from the robot's cameras. For this, the robot would perform an exploratory movement with an object in its hand (such as turning it and looking at it from all angles) and from the resulting image sequence learn spatio-temporal, view-based representations using the framework depicted in Figure 2. Each view of this representation, however, is also linked to the current proprioceptive state (i.e., the joint angles of the hand at that point in time) and therefore provides an anchor into a hand-centered, three-dimensional space. In this way a representation is generated that *links perception and action*. The proprioceptive information can then be used as an additional constraint for both learning of objects and recognition of objects and was shown to provide increased robustness compared to visual matching alone.

64.3 Example-based movement representations

The recognition and modeling of complex movements and actions is fundamental for many applications in robotics. For example, interactive robots need to analyze users' movements in order to respond in a natural way to their behavior. This problem is specifically relevant for the realization of social and emotional behavior in robots (Chapter 59). Another important domain is the recognition and modeling of complex movements for humanoid robots (cf. Chapters 57 and 60) in order to learn new movements by imitation, and to mimic human-like motor behavior.

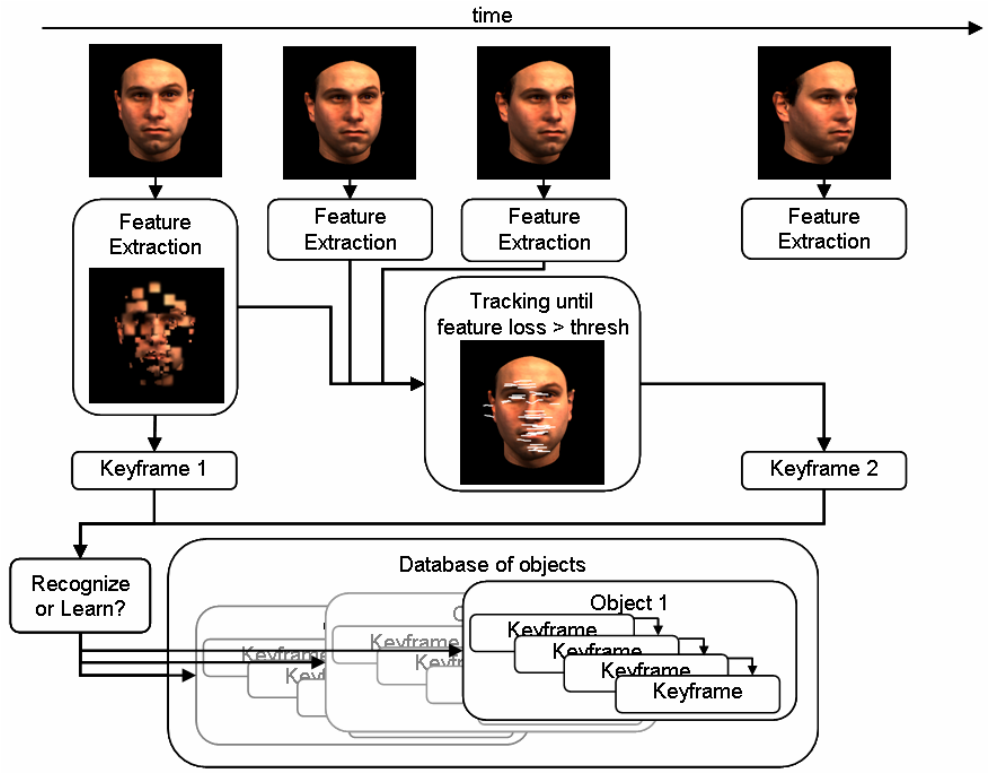

Figure 2. The keyframe representation. This computational framework is a simple architecture which incorporates temporal input, incremental learning, structured object representations as well as efficient indexing. Note that this framework was also used for active exploration coupled with proprioceptive information (see text).

64.3.1 Recognition of complex movements in visual cortex

The recognition of complex movements is also a fundamental problem for biological systems. While simple movement patterns are sufficient for eliciting stereotypical prey catching behavior in simple vertebrates (Ewert, 1970; see also Chapter 63), higher animals exploit more complex movement patters, e.g. for the recognition of conspecifics or predators, or for communication by facial movements, gestures, or body expressions. Human perception of complex motion patterns is very efficient, even for extremely impoverished stimuli. This has been demonstrated in classical experiments by Johansson (1973), who showed that complex dynamic actions can be recognized even from displays that consist only of a small number of dots moving like the joints of a human actor. Subsequent research has demonstrated that humans can extract highly specific information from such *point-light displays*, e.g. the gender or the identity of people. To our knowledge, no technical system for motion recognition has been proposed so far that accomplishes a comparable level of robustness. While much more research in neuroscience has been dedicated to object recognition (Section 64.2.2), some studies have tried to uncover neural (Verfaillie, 2000; O'Toole et al. 2002; Puce & Perrett, 2003) and computational principles (e.g. Hoffman & Flinchbaugh, 1982; Webb & Aggarwal, 1982; Giese & Poggio, 2003) of visual movement recognition. Some of these principles seem interesting for the construction of robotics systems.

Neurophysiological and brain-imaging studies indicate that the recognition of facial and body movements involves the ventral and the dorsal visual pathway. This implies that likely form and optic flow information are integrated during the processing of action stimuli in visual cortex. The ventral pathway, which is specifically responsible for the processing of form information has been discussed already in Section 64.2.2. Like the ventral stream, also the dorsal pathway is hierarchically structured, and the size of the receptive fields of the neurons increases along the hierarchy. Some cortical areas that are part of the dorsal pathway are listed in Figure 3. The *medial temporal area* (MT) contains neurons that are selective for simple local motion and coherent motion. On higher levels of the dorsal stream, e.g. in the *superior temporal sulcus* (STS), neurons that are selective for hand and body movements and for facial expressions have been found in monkeys and humans (Perrett & Puce, 2003). It seems likely that areas selective for human body shapes, e.g. the *extrastriate body part area* (EBA), contribute to the recognition of actions, even though some of them are not specifically selective for motion (e.g. Peelen, Wiggett & Downing, 2006). Finally, neurons that are active during action recognition have also been found in the 'mirror neuron system', which includes regions in the premotor and parietal cortex (e.g. Rizzolatti & Craighero, 2004). Parietal areas might also be specifically relevant for the recognition of action-related objects and their relationship to moving effectors (e.g. Sakata 2003). The mirror neuron system has been very influential in recent research on the construction of biologically inspired robots (see Chapter 63).

A first biological principle of movement recognition that is supported by experimental data, and which seems useful for technical systems, is action recognition by detection of specific relevant motion and form features. Such detection does not necessarily require a reconstruction of facial or body shape, nor an exact simulation of the dynamics of the underlying movements. Instead it can be accomplished by much simpler feature detectors. Like object recognition, the recognition of complex motion patterns is strongly orientation- and view-dependent. This property has also been observed at the level of individual neurons in the STS, and for the activation of biological motion-selective areas in human cortex (Puce & Perrett, 2003). View- and orientation dependence seem compatible with an encoding of visually perceived movements in terms of potentially learned example views or 'keyframes' (Verfaillie, 2000; Giese & Poggio, 2003). While there might be some innate preferences for specific features, psychophysical and fMRI experiments suggest an important role of learning in visual movement recognition (e.g. Grossman et al. 2004; Jastorff et al. 2006). For example, subjects can learn easily to recognize individually–specific body and facial movements (e.g. Hill & Pollick, 2000; Knappmeyer, Thornton & Bülthoff, 2003). Also the fact that learning-based theoretical models, exploiting similar principles as neural object recognition models, can account for a variety of the experimental data (e.g. Giese & Poggio, 2003; Lee & Wong, 2004) supports a central role of learning.

As an example, Figure 3 illustrates a hierarchical model for the recognition of complex movements that has been described in Giese & Poggio (2003). Its neural architecture consists of two hierarchical streams modeling the ventral and the dorsal visual pathways, which contain detectors for action-specific motion and form features. The form pathway of this model is similar to the object recognition models described in section 64.2.2. The motion pathway of the model contains detectors for action-specific optic flow features with different complexity. Like for the described object recognition models, position and scale invariance is accomplished by appropriate nonlinear pooling of the

responses of detectors with different spatial and scale selectivity along the hierarchy. In addition, the model contains recurrent neural circuits that make the responses of the recognition neurons selective for temporal order. In this way the model responds only to actions that are executed with the correct temporal order, and also with approximately correct speed. The underlying network dynamics can be interpreted as a neural implementation of a Markov model, where the present recognized pattern predicts possible future patterns. A strong activity in the network emerges only when the stimulus sequence matches these predictions. A similar neural implementation exploiting the SPIKENET architecture by (Delorme and Thorpe, 2003) (section 64.2.3) has recently been applied in the context of a technical system for tracking of walkers in video sequences (Kornprobst et al. 2005). Hierarchical neural architectures inspired by the visual cortex have also been used in the context of mirror-neuron robots (e.g. Metta et al. 2006).

A second principle of movement recognition that is important for the recognition of *imitable* actions, i.e. actions that can be executed by the observer, is an internal simulation of the observed motor behavior. While the exact nature of the underlying representations has not yet been clarified, a large body of experimental evidence suggests that observation of imitable movements activates the 'mirror neuron system' in premotor and parietal cortex (Rizzolatti & Craighero, 2005). Computational models for action recognition by internal simulation have been proposed, e.g. exploiting feed-forward controllers (e.g. Wolpert et al 2003). Theoretical models for the mirror neuron system with relevance for robotics are discussed in more detail in Chapters 60 and 63.

64.3.2 Example-based motion recognition in robots and computer vision

Many systems for motion recognition and tracking in computer vision demonstrate the computational feasibility and applicability of the first principle: recognition by extraction of action-specific form and motion features without explicit modeling of the geometric shape of bodies or faces (e.g. Bobick, 1997; Essa & Pentland, 1997). In fact, most existing recent systems for face and body tracking are based on this approach. Opposed to the discussed neural models, many of these technical systems exploit non-biological feature detectors and mechanisms for accomplishing temporal order-selectivity and propagation of information over time (e.g. Hidden Markov Models (HMMs), and Kalman or particle filters; Agarwal & Cai, 1999; Gavrila, 1999; Moeslund & Granum, 2001). Even though most of these solutions are not based on neural mechanisms, they highly relevant in robotics for gesture recognition or the tracking of user behavior.

The second principle of action recognition by internal simulation, or by on-line fitting of geometrical or dynamical models, matches the traditional approach in computational vision, which had already been proposed by D. Marr (Marr & Vaina, 1980). A variety of implementations for model-based tracking and recognition of body movements (e.g. Wachter & Nagel, 1999) and facial movements (e.g. Essa & Pentland, 1997) have been proposed in computer vision. Typically, such methods combine 2D or 3D models, and sometimes even dynamic muscle models with predictive models, like Kalman or particle filters (reviews see Agarwal & Cai, 1999; Gavrila, 1999; Moeslund & Granum, 2001; Wang & Singh, 2003).

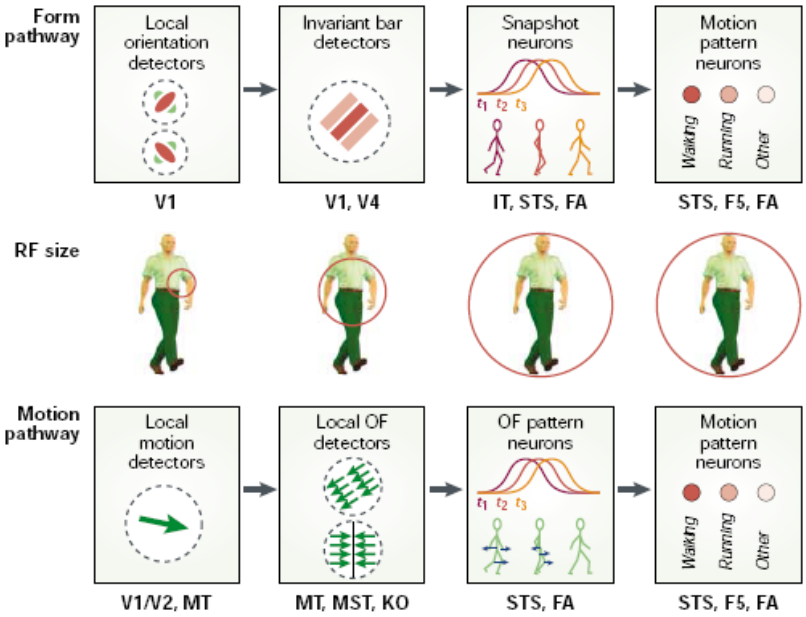

Figure 3. Example-based neural model for the visual recognition of body movements that integrates the processing of form and motion features (from Giese & Poggio, 2003).

While in the biological system the exact nature of the representation of actions in the mirror neuron system is still quite unclear, it seems likely that the brain combines recognition of specific motion and form features and some form of internal simulation of motor programs for the recognition of imitable movements. Likewise, recent approaches in computer vision, computer graphics and robotics try to combine the feature-based recognition (e.g. of body silhouettes) and model-based tracking for the reconstruction of 2D or 3D shape models from video streams (e.g. Rosales & Sclaroff, 2000; Agarwal & Triggs 2006). In fact, the combination of view-based recognition and model-based tracking improves the accuracy and robustness of body tracking (e.g. Curio & Giese, 2005).

63.4. Example-based synthesis models: from faces to movements

In sections 64.2 and 64.3 biologically inspired approaches for the example-based recognition of shapes and movement patterns were discussed. For biological systems it seems likely that the visual recognition of complex patterns and the synthesis of complex patterns in motor control share fundamental principles, like example-based representations that are capable of efficient generalization (Poggio & Bizzi, 2004). Recurring to the most abstract level of analogies between biological and technical systems, example-based models have also been applied for the synthesis of complex patterns in technical applications, such as highly realistic face pictures and movements. These applications are highly relevant for robotics since they form the basis of the simulation of human-like behavior with high believability. The increasing importance humanoid robots (Chapters 57, 59, and 60) and of robotics systems that can interface with normal human social behavior (Chapter 59) makes believable simulation of human behavior a central topic in modern robotics. In several cases, example-based learning techniques have been superior to classical methods based on geometrical or physical models for such applications. A further application domain for example-

based synthesis is the automatic generation of virtual training examples for learning-based robotics (Chapter 60) or computer vision systems (e.g. Niyogi et al. 1998).

63.4.1 Face synthesis: realistic faces for avatars and robots

One of the crucial ingredients of careful studies of face processing – both for facial expressions and identity recognition – is a highly controllable, yet at the same time realistic stimulus set. While it is certainly possible to do research with impoverished or highly abstracted stimulus sets (such as line drawings) or only in the static domain (the majority of research has been done using static photographs), the results are certainly limited in terms of their generalizability to the real-world. One of the recent thrusts in cognitive research has therefore been to use highly realistic face renderings – both static and dynamic – using advanced computer graphics techniques. As this technology offers novel insights also in the context of robotics, we will briefly describe a high-end facial animation system that was used in several studies. One important application domain of synthetic faces is for communication and interaction in any scenario in which communication between a robot and a human is required. Being able to synthesize realistic faces allows the robot to communicate its internal state and intentions in a manner that is very natural and easy to understand for humans. In addition, this research can be used not only for synthesis purposes but also for analysis purposes – that is to design vision systems for robots that understand and react to the complexity of human faces appropriately. Again, we envision such a recognition and analysis system to be an essential component for any robotic application that relies on interactions with humans – especially in the field of anthropomorphic robotics.

The idea of a face space – that faces are represented in a high-dimensional vector space – has been pioneered by Valentine in 1991 and has been influential especially in perceptual research on face processing. A number of techniques in computer vision and computer graphics have implemented the concept of a face space, first by interpolation or morphing between example pictures (e.g. Vetter & Poggio, 1997; Lanitis et al, 1997). As further step toward the simulation of highly realistic pictures of faces, Blanz and Vetter (1999) created a three-dimensional face space from 3D laser scans of faces. The critical step for defining a vector space of faces was to bring the example patterns into correspondence. For 3D laser scans this means that, for example, the tip of the nose can be uniquely identified in all scans. By using a modified optic flow algorithm, which establishes a dense correspondence between two 3D scans, Blanz and Vetter were able to create such a face space. Such vector space models generated by interpolation between example pictures (or scans) that are in spatial correspondence are also called *morphable models* (see Figure 4). Based on such models the properties of face images, e.g. their male- or femaleness, can be continuously varied. This allows the synthesis of artificial new facial expressions, or of face pictures with novel identities. Also existing pictures of faces can be modified, resulting in novel pictures with photorealistic quality (Blanz & Vetter, 1999).

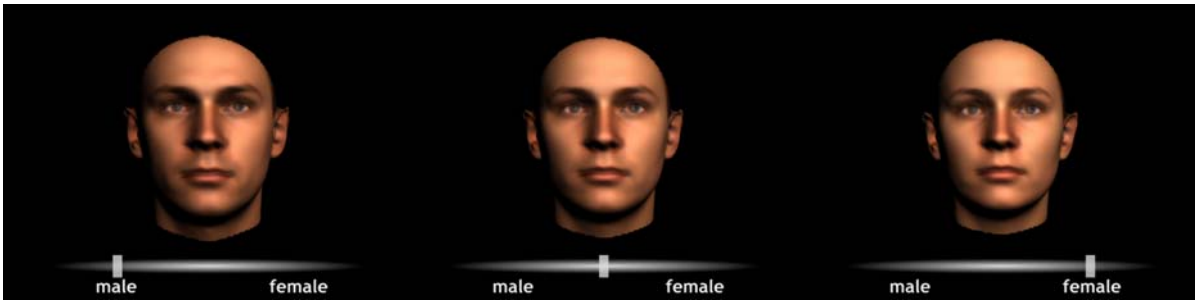

Figure 4. Morphable model: The three images show a morph from an average male face to an average androgynous face to an average female face.

By analysis of the statistical properties of face spaces, perception researchers have been able to conduct highly controlled experiments, e.g. addessing gender perception (Bülthoff & Newell, 2004) or identity coding (O'Toole et al., 1999). In addition, an encoding of face stimuli in terms of a vector space seems to be compatible with recent psychophysical results in humans (Leopold et al., 2001), and electrophysiological results in monkeys (Leopold et al., 2006). This makes this concept interesting as basis for the design of perceptual robots that have to synthesize and analyze faces (Chapter 59).

Recently, static morphable models were extended as the basis of a dynamic animation system (Breidt et al., 2003). The basic idea is to combine the high spatial resolution of 3D laser-scans with the high temporal resolution of motion capture. Facial geometry is first captured using a structured light scanner, which produces a detailed 3D-scan of each peak expression. The individual scans are put into correspondence with each other based on a control mesh whose 3D fiducial points are aligned to each scan which results in a morphable model in much the same way as the one described earlier. The second step in creating the animation system is to capture motion data with an optical motion capture system. On this motion capture data, simple linear detectors for facial action elements are specified by using distances between markers to drive the expression activation. This yields a realistic animation based on the amplitude and timing of marker motion in the motion capture data. The morph animations of the final animation system thus allow full control over important information channels used in facial expressions (such as internal motion of the face, rigid head motion, or fidelity of shape and texture) making it ideally suited for experiments on perception of facial expressions (Wallraven et al., 2005). The results from such experiments could, for example, be used in perceptual robotics to design robots that are able to understand (being able to analyze what is said and meant by a human) and react (being able to synthesize a realistic facial expression) to the variety of human facial expressions in a realistic fashion thus enabling life-like interaction with users.

63.4.2 Example-based trajectory synthesis

Example-based techniques have been successful for the synthesis of photorealistic pictures (64.4.1.) and the recognition of complex movements (64.3.1.). It turns out that related methods are also suitable for the modeling of complex movement trajectories by interpolation between learned examples. This approach has been very successful in computer graphics, specifically for the modeling of highly realistic human movements. The simulation of realistic

human movements for platforms with many degrees of freedom, like avatars and humanoid robots, based on physical or dynamical models is extremely difficult. Learning-based techniques exploiting motion capture data have thus become the most successful approach for the high-quality synthesis of complex human movements. A variety of methods for the synthesis of movement trajectories by interpolation and editing of motion-captured trajectory segments have been developed in computer graphics. Many of these approaches are also interesting for movement synthesis in robotics.

Simple approaches for example-based trajectory modeling by interpolation define novel movement styles by *motion morphing* between pre-segmented trajectories (Unuma et al. 1995; Bruderlin & Williams, 1995). An example is the interpolation between different gait styles. A variety of methods have been proposed for motion morphing (review see Wiley & Hahn, 1997), i.e. the blending between different movements styles. Some of these methods define linear Euclidean style spaces by linear combination of example trajectories (e.g. Yacoob & Black, 1999; Giese & Poggio, 2000; Safonova et al., 2004). Such style spaces can be seen in analogy to face spaces. Linear combination methods for trajectories represent thus a spatio-temporal equivalent of morphable models (discussed in Section 64.4.1). Another class of methods interpolate between example trajectories by learning of generative statistical models for the space-time texture of the training patterns. Such methods often yield real-time capable stochastic models for the generation of novel trajectories, e.g. exploiting Hidden Markov Models (e.g. Brand & Hertzman, 2000). Real-time capability is essential for robotics applications, since usually it is not possible to pre-synthesize complete movements off-line. Instead, robotic systems need to be reactive and capable of an on-line correction of synthesized movements depending on changes in the environment.

While simple motion interpolation techniques require pre-segmented and aligned example trajectories, more advanced techniques have been proposed that accomplish an automatic segmentation of complex action streams into simpler components, or *movement primitives*. The concept of movement primitives has been quite influential in the theory of motor control of biological systems. A variety of different definitions of primitives has been given in the literature of motor control, ranging from coordinated activation patterns (synergies) at muscle level to more abstract cognitive units of movement planning (review Flash & Hochner, 2005). In the context of robotics different approaches have been taken to identify and implement movement primitives. Sometimes, primitives have been predefined heuristically, e.g. as single goal-directed or periodic movements. Recent studies have tried to learn primitives directly from trajectory data sets applying supervised or unsupervised learning techniques, for example PCA or ISOMAP (e.g. Jenkins & Mataric, 2004; Safonova et al. 2004). Temporal correspondence between learned primitives and movements novel sequences can be established by appropriate dynamic alignment algorithms (e.g. Ilg et al. 2004). Optimized transitions between pre-stored or synthesized trajectory segments can be generated automatically, e.g. exploiting graph optimization algorithms (e.g. Arikan & Forsyth, 2003; Kovar & Gleicher, 2002). In this way, trajectories for complex action sequences can be synthesized automatically from learned example trajectories, ensuring that they fulfill specific style properties and are consistent with additional kinematic or dynamic constraints.

Some studies have embedded such learned trajectory models into real-time capable systems, e.g. for on-line tracking and synthesis of movement trajectories interactive applications like computer games (Yacoob & Black,

1999; Ren et al. 2005). Such methods seem highly suitable for online imitation of movements in humanoid robots. Finally, movement primitives can also be modeled as dynamical control policies. It has been shown that such policies can be encoded by nonlinear dynamical systems whose stable solutions reproduce complex training movements (Schaal et al. 2003) (see Chapter 60 for details).

Movements generated by interpolation between example trajectories can violate constraints of robot systems, e.g. kinematic or dynamic limits of the platform and the environment. One possibility is to use the synthesized movement as *virtual trajectory* that specifies a temporal sequence of goal states for the robot controller (e.g. Ilg et al. 2004). A variety of 'filtering techniques' have been proposed that deform the synthesized trajectories in a way that makes them compatible with additional spatial or temporal constraints (e.g. Gleicher & Litwinowicz, 1998). By exploiting physical models also compatibility with dynamic constraints, like the minimization of torques, can be accomplished (e.g. Safonova et al. 2004). For walking avatars and robots additional dynamic stability constraints have to be taken into account, e.g. concerning the dynamic stability of walking. Methods for deforming synthesized trajectories to make the compatible, for example with the Zero Moment Point (ZMP) criterion (Chapter 16), have been proposed (e.g. Tak & Ko, 2005). Such approaches try to combine the advantages of example-based trajectory generation (high degrees of accuracy and realism) with the ones of classical physics-based trajectory planning (real-time capability, simplicity of expressing physical and task constraints). Yet, the optimal integration of efficient learning models of complex movements and classical concepts in robot control is an active field of research in learning-based and humanoid robotics (cf. Chapter 60).

63.5. Summary / Outlook

In this chapter we have presented several principles derived from high-level cognitive processing in vision in the human brain that have been fruitful for the development of systems in robotics and computer vision. The recognition of shapes and complex movements and actions is an important problem for many applications in robotics. We have discussed a variety of results from neuroscience that indicate that the brain these functions are likely realized by example-based representations. We have discussed neural implementations of such representations which have been tested in the context of technical applications, and which are strongly inspired by the real cortical neural architecture. In addition, we have discussed a class of methods for the highly accurate modeling of complex shapes and movement trajectories that are also based on learned examples, but which exploit more efficient technical algorithms for interpolation between the example patterns. We finally have discussed also approaches that try to link such example-based representations with more traditional model-based recognition and tracking approaches. In addition, we have mentioned some methods for adapting patterns that are synthesized by such example-.based techniques to boundary conditions and constraints of real robot platforms, and to accomplish real-time capability. We think that the discussed spectrum of results and applications shows that a transfer between biological and robotics systems on multiple levels can help to develop more efficient technical systems.

Example-based mechanisms for object and motion recognition account for the invariant recognition of complex patterns: However, they do not automatically extract the metric information about the object geometry, position and the spatial parameters of complex trajectories in world coordinates. For some tasks in robotics, like grasping,

manipulation or obstacle avoidance, such information is required (e.g. Chapters 28 + 52). For such tasks, example-based recognition must be fused with methods for the extraction of the relevant metric information. In robotics such information can be extracted by stereo vision or using special sensors, like laser range finders. In the brain the fusion between such spatial information and information about objects occurs likely in parietal areas, like the *anterior interparietal area* (AIP) (Colby, 1998). However, it is unclear whether the information about objects is only represented in terms of 2D example views. Instead, it seems likely that also some form of 3D information is encoded, potentially in an example-based manner. Also haptic and visual information about object shape might be merged in higher brain areas, e.g. in parietal and fusiform areas (e.g. Kilgour et al. 2005). A more detailed discussion about biologically inspired models for the extraction of action-relevant geometrical information of objects and its combination with spatial information for grasping is discussed in the context of the mirror neuron system (Chapter 63).

In section 64.2.1 we have contrasted example-based and model-based theories of the representation of objects in the brain. In fact, in a recent paper an attempt has been made at reconciling these two approaches to object processing (Foster and Gilson, 2002): a careful study of the view-dependency of novel objects was designed by combining structural properties (number of parts) with metric properties (thickness, size of parts) has found that both view-dependent and view-independent processing seem to be combined in object recognition. Thus, instead of taking the extreme standpoints of view-based versus view-invariant processing one might envisage a visual processing framework in which features are selected according to the current task, where the optimality, efficiency and thus the dependency on viewing parameters of the features depend on the amount of visual experience with this particular task. This suggests that a similar flexible approach might be suitable for the solution of problems in robotics. The combination of metric model-based representations for the planning of goal-directed movements and of example-based representations for the recognition of objects, people, and actions might result in powerful robotics systems with performance levels more similar to biological systems than the presently available solutions in robotics.

64.6. References

A. Agarwal, and B. Triggs. Recovering 3D Human Pose from Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),* 28, 44-58, 2006.

J.K Aggarwal, and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding,* 73:428-440, 1999.

E. Ashbridge and D.I. Perrett. Generalizing across object orientation and size. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancy. Why things look as they do* (192-209). Cambridge: Cambridge University Press, 1998.

O. Arikan, and D.A. Forsyth, and J.F. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics*, 22, 402-408, 2003.

I. Biederman. Recognition-by-components: a theory of human image understanding. Psychological Review, 94, 115-147, 1987.

V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *In Proc. ACM SIGGRAPH*, 187–194, 1999.

A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B* 352, 1257-1265, 1997.

M. Brand, and A. Hertzmann. Style machines. *Proceedings of the ACM SIGGRAPH 2000,* ACM Press, New York, 183-192, 2000.

M. Breidt, C. Wallraven, D.W. Cunningham and H.H. Bülthoff. Facial Animation Based on 3D Scans and Motion Capture. SIGGRAPH '03 Sketches & Applications., 2003.

A. Bruderlin, and L. Williams. Motion signal processing. *ACM Transactions on Computer Graphics*, 29:97-104, 1995.

H.H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 60-64, 1992.

I. Bülthoff and F.N. Newell. Categorical perception of sex occurs in familiar but not unfamiliar faces. *Visual Cognition,* 11, 823-855, 2004.

K.R. Cave, S. Pinker, L. Giorgi, C.E. Thomas, L.M. Heller, J.M. Wolfe and H. Lin. The representation of location in visual images. *Cognitive Psychology*, 26, 1-32, 1994.

C.L. Colby. (1998) Action-oriented spatial reference frames in cortex. *Neuron*, 20,15-24.

C. Curio, and M.A. Giese. Combining view-based and model-based tracking of articulated human movements. *IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION 2005)*, IEEE Computer Society Press, Los Alamitos, 261-268, 2005.

A. Delorme and S. Thorpe. SpikeNET: An Event-driven Simulation Package for Modeling Large Networks of Spiking Neurons, *Network: Comput. Neural Syst.*, 14, 613-627, 2003.

M. Dill and S. Edelman. Imperfect invariance to object translation in the discrimination of complex shapes. *Perception*, 30, 707-724, 2001.

S. Edelman. *Representation and Recognition in Vision*. Cambridge, MA, MIT Press, 1999.

I. Essa, and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 757-763, 1997.

J.-P. Ewert. Neural mechanisms of prey-catching and avoidance behavior in the toad (bufo bufo). *Brain, Behavior, and Evolution* 3, 36-56, 1970.

D.H. Foster and S.J. Gilson. Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society of London B* 269, 1939–1947, 2002.

D.M. Gavrila. The visual analysis of human movement: a survey. *Comp. Vis. Image Underst*. 73, 82-98, 1999.

M.A. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *Int. J. Comp.* Vis. 38, 59-73, 2000.

M.A. Giese and T. Poggio, T. Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179-192, 2003.

M. Gleicher and P. Litwinowicz. Constraint-based motion adaptation. *Journal of Visualization and Computer Animation* 9, 65-94, 1998.

M. Graf and W. Schneider. Structural descriptions in HIT - a problematic commitment. *Behavioral and Brain Sciences,* 24, 483-484, 2001.

K. Grill-Spector and R. Malach. The Human Visual Cortex. *Annual Review of Neuroscience*, 27, 649 -677, 2004.

E.D. Grossman, R. Blake, and C.Y. Kim. Learning to see biological motion: brain activity parallels behavior. *J. Cogn Neurosci.* 16, 1669-1679, 2004.

W.G. Hayward and M.J. Tarr Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1511-1521, 1997.

H. Hill and F.E. Pollick. Exaggerating temporal differences enhances recognition of individuals from point light displays. *Psychological Science*, 11, 223-228, 2000.

H. Hill, P.G. Schyns and S. Akamatsu. Information and viewpoint dependence in face recognition. *Cognition* 62, 201-222, 1997.

D.D. Hoffman, and B.E. Flinchbaugh. *The interpretation* of biological motion. *Biological Cybernetics*, 42, 195-204, 1982.

D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160, 106–54, 1962.

W. Ilg, G.H. Bakir, J. Mezger and M.A. Giese. (2004) On the representation, learning and transfer of spatio-temporal movement characteristics. *International Journal of Humanoid Robotics* 1, 613-636, 2004.

J. Jastorff, Z. Kourtzi and M.A. Giese. Learning to discriminate complex movements: Biological versus artificial trajectories. *Journal of Vision,* 6, 791–804, 2006.

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

G. Johansson. (1973) Visual perception of biological motion and a model for its analysis. *Perc. Psychophys*. 14, 201-211.

J.A.S. Kelso. Dynamic Patterns: The Self-Organization of Brain and Behaviour, MIT Press, Cambridge, MA, 1995.

A.R. Kilgour, R. Kitada, P. Servos, T.W. James, S.J. Lederman. Haptic face identification activates ventral occipital and temporal areas: an fMRI study. *Brain Cogn.* 59:246-257, 205.

M. Kirby and L. Sirovich. Applications of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern and Machine Intelligence* 12, 103–108, 1990.

B. Knappmeyer, I.M. Thornton, and H.H. Bülthoff. The use of facial motion and facial form during the processing of identity. *Vision Res.* 43, 1921-1936, 2003.

P. Kornprobst, T. Vieille, I.K. Dimo. Could early visual processes be sufficient to label motions? *Proceedings of the IEEE International Joint Conference of Neural Networks (IJNN'05),* 1687-1692, 2005.

L. Kovar, M. Gleicher and F. Pighin. Motion graphs. *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press, New York, 473-482, 2002.

M.A. Kurbat. Structural description theories: is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception* 23, 1339-1368, 1994.

A. Lanitis, C.J. Taylor and T.F. Cootes. Automatic Interpretation and Coding of Face Images Using Flexible Models. *IEEE Transactions on Pattern and Machine Intelligence* 19, 743-756, 1997.

R. Lawson, and G.W. Humphreys. View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. *Perception & Psychophysics*, 60, 1052-1066, 1998.

Y. LeCun, F. Huang and L. Bottou, Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting in *Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2004.

D.A. Leopold, A.J. O´Toole, T. Vetter and V. Blanz: Prototype-referenced shape encoding revealed by high-level after effects. *Nature Neuroscience* 4, 89-94, 2001.

J. Lee, and W. Wong. A stochastic model of coherent motion detection. *Biological Cybernetics*, 91, 306-314, 2004.

D.A. Leopold, I.V. Bondar, and M.A. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442:572-575, 2006.

N.K. Logothetis, J. Pauls, H.H. Bülthoff and T. Poggio: View-dependent object recognition by monkeys. *Current Biology* 4, 401-414, 1994.

D. Lowe. *Perceptual Organization and Visual Recognition.* Boston, MA: Kluwer, 1985.

D. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

D. Marr. *Vision*. San Francisco: Freeman Publishers, 1982.

D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B* 200, 269-294, 1978.

D. Marr, and L.M. Vaina. Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B* 214: 501-524, 1982.

G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga. Understanding mirror neurons: a bio-robotic approach. *Interaction Studies,* 7, in press.

T.B. Moeslund, and G. Granum. A survey of computer vision-based human motion capture. *Comp. Vis. Image Underst*. 81, 231-268, 2001.

F.N. Newell, M.O. Ernst, B.S. Tjan, and H.H. Bülthoff. Viewpoint dependence in visual and haptic object recognition. *Psychological Science* 12, 37–42, 2001.

P. Niyogi, F. Girosi and T. Poggio. Incorporating prior information in machine learnig by creating virtual examples. *Proceedings of the IEEE* 86, 2196-2209, 1998.

A.J. O'Toole, T. Vetter and V. Blanz. Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Research*, 39, 3145-3155, 1999.

A.J. O'Toole, D.A.. Roark, H. Abdi, H. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, 6:261-266, 2002.

S.E. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (135-151). Hillsdale, N.J.: Erlbaum, 1981.

M. Peelen, A. Wiggett, and P. Downing, P. Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron* 49, 815-822, 2006.

D. Perrett and W.M. Oram. Visual recognition based on temporal cortex cells: viewer-centred processing of pattern configurations. *Zeitschrift für Naturforschung, C*, 53, 518-541, 1998.

D. Perrett, W.M. Oram, and E. Ashbridge. Evidence accumulation in cell populations responsive to faces: an account of generalization of recognition without mental transformations. *Cognition*, 67, 111-145, 1998.

D. Perrett, and A. Puce. Electrophysiology and brain imaging of biological motion. *Phil. Trans. R. Soc. Lond. B* 358, 435–445, 2003.

T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature* 431, 768-774, 2004.

T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343, 263-266, 1990.

L. Ren, G. Shakhnarovich, J.K. Hodgins, H. Pfister, and P. Viola. Learning silhouette features for control of human motion. A*CM Transactions on Graphics* 24,1303-1331, 2005.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2, 1019-1025, 1999.

G. Rizzolatti, and L. Craighero. The mirror-neuron system. *Annual Reviews of Neuroscience*, 27, 169-192, 2004.

L. Roberts. Machine perception of three-dimensional solids In: Tippet et al., eds. *Optical and Electro-optical Information Processing*, MIT Press, 159–197, 1965.

E.T. Rolls and T. Milward. A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, 2(11), 2547–72, 2000.

R. Rosales, and S. Sclaroff, S. Inferring body pose without tracking body parts. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume II, 721–727, 2000.

A. Safonova, J.K. Hodgins, N.S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Computer Graphics* 23, 514-521, 2004.

H. Sakata (2003) The role of the parietal cortex in grasping. *Adv Neurol.* 93, 121-139

S. Schaal, A. Ijspeert, and A.. Billard. Computational approaches to motor learning by imitation. *Philosophical Transaction of the Royal Society of London B* 358, 537-547, 2003.

C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern and Machine Intelligence* 19, 530–535, 1997.

G. Schöner, M. Dose, and C. Engels. Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16, 213-245, 1997.

T. Serre, L. Wolf and T. Poggio. Object recognition with features inspired by visual cortex. *In: Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2005.

J.V. Stone. Object recognition using spatio-temporal signatures. *Vision Research*, 947–951, 1998.

J.V. Stone. Object recognition: View-specificity and motion-specificity. *Vision Research*, 4032–4044, 1999.

M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision* 7, 11–32, 1991.

K. Tanaka. Representation of visual feature objects in the inferotemporal cortex. *Neural Networks*, 9(8), 1459–75, 1996.

S. Tak, and H.S. Ko. A physically based motion retargeting filter. *ACM Transactions on Graphics* 24, 98-117, 2005.

J. Tani, and M. Ito. Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment. *IEEE Trans. on Systems, Man, and Cybernetics Part A: Systems and Humans* 33(4), 481-488, 2003.

M.J. Tarr. Rotating objects to recognize them: a case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review* 2, 55-82, 1995.

M.J. Tarr and S. Pinker. Mental orientation and orientation-dependence in shape recognition. *Cognitive Psychology* 21, 233-282, 1989.

S. Thorpe, D. Fize and C. Marlot. Speed of processing in the human visual system. *Nature* 381(6582), 520-522, 1996.

S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32, 193-254, 1989.

S. Ullman. *High-level vision. Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press, 1996.

S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 992-1006, 1991.

M. Unuma, K. Anjyo, R. Takeuchi. Fourier principles for emotion-based human figure animation. *ACM Transactions on Computer Graphics* 29, 91-99, 1995.

T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 43, 161 – 204, 1991.

K. Verfaillie. Perceiving human locomotion: priming effects in direction discrimination. *Brain Cogn.* 44:192-213, 2000.

T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 733-742, 1997.

Q.C. Voung and M.J. Tarr. Rotation direction affects object recognition. *Vision Research* 44*(14)*, 1717-1730, 2004.

S. Wachter, H.-H. Nagel. Tracking of persons in monocular image sequences. *Comp. Vis. Image Underst*. 74,174-192, 1999.

G. Wallis and H.H. Bülthoff. Learning to recognize objects. *Trends in Cognitive Sciences* 3, 22-31, 1999.

G. Wallis and H.H. Bülthoff. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America* 98, 4800–4804, 2001.

C. Wallraven and H.H. Bülthoff. Automatic acquisition of exemplar-based representations for recognition from image sequences. *CVPR 2001 - Workshop on Models vs. Exemplars*, 2001.

C. Wallraven, M. Breidt, D.W. Cunningham, and H.H. Bülthoff. Psychophysical evaluation of animated facial expressions. *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, 17–24, 2005.

C. Wallraven, B. Caputo and A.B.A. Graf. Recognition with Local Features: the Kernel Recipe. *Proceedings of International Conference on Computer Vision 2*, 257-264, IEEE Press, 2003.

C. Wallraven and H. H. Bülthoff: Object Recognition in Man and Machine. *Object Recognition, Attention and Action*, Springer, Tokyo (in press)

C. Wallraven, A. Schwaninger, S. Schuhmacher and H.H. Bülthoff. View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho. *Lecture Notes in Computer Science* 2525, 651-660, 2002.

G. Wang, M.Tanifuji and K. Tanaka. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research* 32, 33-46, 1998.

J.J. Wang, S. Singh. Video analysis of human dynamics - a survey. *Real-Time Imaging* 9, 321-346, 2003.

W.H. Warren. The dynamics of perception and action. *Psychological Review* 113, 358–389, 2006.

J.A. Webb, and J.K. Aggarwal. (1982) Structure from motion of rigid and jointed objects. *Artif. Intell*. 19, 107-130, 1982.

D.J. Wiley, and J.K. Hahn. Interpolation synthesis of articulated figure motion. *IEEE Computer Graphics and Applications* 17, 39-45, 1997.

D.M. Wolpert, K. Doya, and M. Kawato. A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci*. 358, 593-602, 2003.

Y. Yacoob, and M.J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding* 73:232-247, 1999.