

Multimedia Contents



78. Perceptual Robotics

Heinrich Bülthoff, Christian Wallraven, Martin A. Giese

Robots that share their environment with humans need to be able to recognize and manipulate objects and users, perform complex navigation tasks, and interpret and react to human emotional and communicative gestures. In all of these perceptual capabilities, the human brain, however, is still far ahead of robotic systems. Hence, taking clues from the way the human brain solves such complex perceptual tasks will help to design better robots. Similarly, once a robot interacts with humans, its behaviors and reactions will be judged by humans – movements of the robot, for example, should be fluid and graceful, and it should not evoke an *eerie* feeling when interacting with a user. In this chapter, we present Perceptual Robotics as the field of robotics that takes inspiration from perception research and neuroscience to, first, build better perceptual capabilities into robotic systems and, second, to validate the perceptual impact of robotic systems on the user.

78.1 Perceptual Mechanisms of Object Representations 2097

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 (Chaps. 7, 8, 32, 33, 36–38, 47, 67). Social interactive robots must be able to interpret gestures, actions, and even emotions (Chap. 69, 71, 72) in order to interact naturally with their users. One important approach for the programming of complex perceptual and behavioral functions, for example, needed for humanoid robots is imitation learning (Chaps. 75, 77). Imitation learning requires the robot to *perceive* complex actions

78.1.1	Perceptual and Computational Basis of Object Representations ..	2097
78.1.2	Neural Representations in Object Recognition	2100
78.1.3	Object Recognition: Lessons from Computer Vision	2101
78.1.4	Object Learning and Recognition for Perceptual Robotics	2102
78.2	Perceptual Mechanisms of Action Representation	2103
78.2.1	Recognition of Complex Movements and Actions in Primate Cortex	2103
78.2.2	Biological Principles with Relevance for Computer Vision and Robotics	2104
78.3	Perceptual Validation of Robotics	2107
78.3.1	Realistic Faces for Robots	2107
78.3.2	Perceptual and Neural Processing of Body Movements of Robots	2108
78.4	Conclusion and Further Reading	2108
	Video-References	2109
	References	2109

that are executed by the user and to subsequently map them 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 (Chap. 41), auditory perception, sensory cue fusion (Chap. 35), and the interaction between the visual recognition of objects and actions and motor programs, e.g., during grasping (treated in Chap. 38). 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 principles 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 [78.1]: 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 Chap. 77. 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 3-D structure, e.g., using a parametric 3-D 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 the cases, the robot system realizes mechanisms that are derived from 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 and behavior-based approaches (Chap. 13) that decompose robotics system into a system of simpler behavioral modules. Another examples are dynamical systems approaches to robotics [78.2–4] 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 [78.5].

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 us 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 (Chaps. 13, 71, 74, 75). On the other hand, perception science often uses robots as testbed for gaining a deeper understanding of computational processes, 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.

Finally, perceptual robotics not only means to take inspiration from perception to build more efficient robots, but it also encapsulates the perceptual validation of robotic systems. As robots move into the human environment and are increasingly also interacting with

humans, it becomes important to evaluate and validate their effectiveness and efficacy with respect to human standards. Here, we do not refer to their social acceptance, but rather to the way that robots are judged by humans in terms of their appearance, movements, and interactive capabilities. If a robot displays jerky movements, for example, it may still successfully grasp and manipulate an object, but it would be immediately noticeable to a human observer and potentially disturbing to interact with. This *eeriness* or *weirdness* was already anticipated in the early 1970s in a famous paper about the *uncanny valley* by a Japanese roboticist [78.6]. Mori anticipated that as robots become more human-like, humans' familiarity with the robot would increase until at some point (when the robot looks or acts almost human-like), they would suddenly feel highly unfamiliar toward the robot. As the human likeness increases further, the robot would again be judged as familiar or appealing.

More specifically, Mori also postulated that this uncanny valley would not only hold for the robot's static appearance, but would in fact be increased for a moving or acting robot. With the increase in interest in develop-

ing humanoid robots over the past decades, being aware of the perceptual judgments of such humanoids becomes a critical component in their development. Since the evaluation of appearance and movements of a humanoid are driven by perceptual processes, it makes sense to also use protocols from perception research to evaluate and fine-tune their effectiveness. In such experiments, typically the robot's performance is evaluated with respect to measures such as general user acceptance, recognizability of expressions, smoothness of motions, ease of interaction, duration and quality of interaction, etc. It is important that the experiment should not only be about simply asking *how good is the robot*, but it should actually tests the robot in the intended task context or that whether it uses additional, indirect measures of effectiveness. As a tutorial on designing and analyzing perceptual experiments and user studies is beyond the scope of this chapter, we refer the reader to introductory texts such as [78.7, 8]. In this chapter, we will focus on two important topics related to humanoid perception in the context of perceptual robotics: facial animation and the perceptual processing of body movements.

78.1 Perceptual Mechanisms of Object Representations

Object recognition is a fundamental visual function that is critical for many applications in robotics. Manipulation and grasping (Chaps. 36–38) require exact knowledge about the shape of the goal object that is often derived from visual sensors. Also the imitation of goal-directed movements (Chap. 77) 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 (Chaps. 71, 72). The importance of object and shape recognition for many other applied robot systems, like construction and assembly robots or smart cars (Chap. 54) is immediately evident.

78.1.1 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 geomet-

ric primitives (cf. also Chap. 32). These primitives are then matched to an internal, three-dimensional model of the object (Fig. 78.1, 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 categorization is based on a structural representation, which is defined as a configuration of elementary object parts that are regarded as shape primitives [78.9]. 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 (RBCs) or geon structural description (GSD) model can be regarded as the best developed example of the struc-

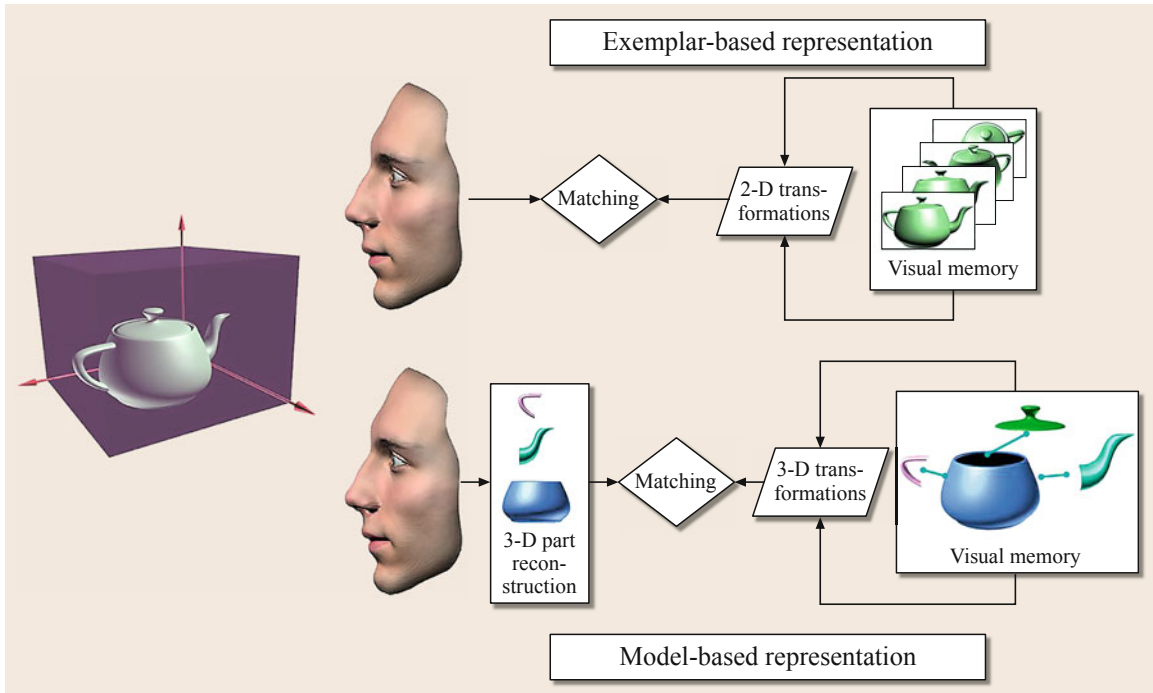


Fig. 78.1 Schematic drawing comparing exemplar-based with model-based representations. Object perception based on model-based representations assumes that the brain extracts 3-D parts from the visual image, which are then matched to an internally stored 3-D model of the teapot. Contrasting with this 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

tural description model type [78.10]. 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 NAPs (NAPs were originally proposed within an image-based approach by Lowe [78.11]). According to the model, geons and their spatial configuration are combined into a structural representation, called GSD. 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.

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 [78.12], or that biological shapes in general cannot be adequately described by struc-

tural description models [78.13]. This problem extends also to artifact 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 (further problems of RBC in [78.14, 15]). However, this does not mean that category representations do not have a part structure: in fact, it is not the notion of the 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 [78.15].

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 [78.16, 17], and also for common, familiar objects [78.18, 19]. Orientation-dependent recognition performance has been shown not to be limited to individual objects, such as faces [78.20, 21], or to objects on the subordinate level of categorization [78.16, 22], but also was demonstrated for basic level recognition [78.19, 23].

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 *Ashbridge and Perrett [78.24]*). 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 [78.25] and familiar objects [78.26]. 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 [78.12, 27]* 3-D alignment model and *Lowe's [78.11]* SCERPO model are probably the best-known examples. Both models work by storing 3-D 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 [78.27]* model that relies on 3-D object representations, *Ullman and Basri [78.28]* suggested an alignment model on the basis of 2-D (two-dimensional) views. In this model, an internal object model is constructed by a linear combination of a small number of stored 2-D 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 3-D model.

In the interpolation model, recognition is achieved by localization in a multidimensional representational space, which is spanned by stored views [78.29]. The interpolation model is based on the theory of approximation of multivariate functions and can be imple-

mented 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 1990s – and as an extension to the interpolation models – recognition models based on pooling and thresholding were developed [78.30–33]. 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 [78.30]. A similar proposal was made by *Riesenhuber and Poggio [78.31]*. The threshold model [78.34] 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-dependence without the need to postulate transformation or interpolation processes.

In a recent paper, an attempt has been made at view-dependent and view-independent approaches to object processing [78.35]. 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.

Several computational models have been proposed that aim at modelling and explaining the dependence of human recognition performance on spatial transformations in its complexity. All of these models rely on storing exemplars – in the simplest form just 2-D 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.

78.1.2 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 [78.36] 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 (Chap. 77 for a more detailed discussion).

The seminal work of *Hubel and Wiesel* [78.37] 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 also 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 ([78.38]; see also *Tanaka* [78.39] 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.* [78.38] 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 ([78.40] for a review).

In another set of experiments, *Logothetis et al.* [78.41] found AIT neurons, which showed a strong view-based behavior for the same stimuli that were used in the study of *Bülthoff and Edelman* [78.16], 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 [78.33].

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 feedforward fashion, feature complexity increases from simple edge detectors toward view-tuned, com-

plex 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.

78.1.3 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* [78.42] 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 3-D reconstruction [78.1]. 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 3-D 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 [78.43], local feature detectors [78.44, 45] or on a straightforward pixel representation of images using principle components analysis [78.46]. 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 neuromorphic recognition systems that are based on a functionally plausible, exemplar-based architecture: these are SpikeNET [78.48], LeNet [78.49], and a framework by *Serre et al.* [78.50]. The first system is motivated by the finding that humans are amazingly fast at categorizing images as containing an animal or

a face [78.51]. Typical response times for this task are so small (on the order of 100 ms) that the visual signal has only time for one feedforward pass through the visual areas of the brain (Fig. 78.2) – 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 [78.48] 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 [78.37]. This spike time encoding allows for very fast processing of visual stimuli and has been shown to provide robust recognition results. The network ar-

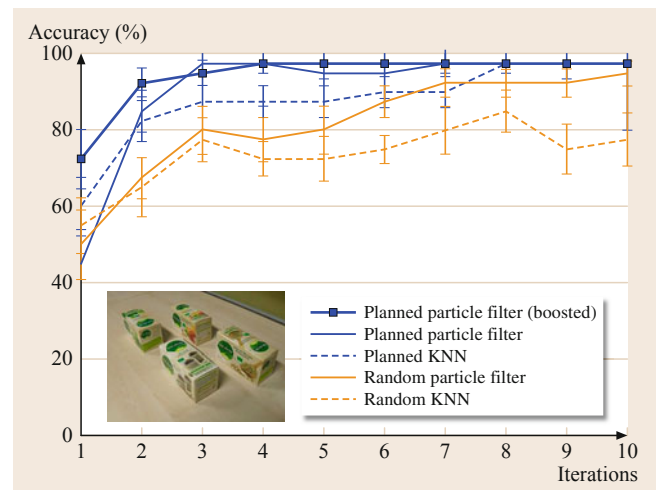


Fig. 78.2 Recognition performance of four highly similar object (shown in the inset) by an in-hand recognition system using active view selection (after [78.47]). The five methods compared in the plot contrast planned (*blue*) and unplanned (*orange*) exploration of the objects in the hand of the robot. The *x*-axis is the number of iterations, and the *y*-axis is the recognition accuracy in percent. As time (or iteration number) proceeds, the planned approaches surpass random exploration significantly. In addition, employing probabilistic methods for recognition of the objects using a particle filter also provides a recognition improvement. Finally, the *thick, solid blue line* shows performance in a system which boosts the likelihood of an object given the current visual evidence in the particle filter framework – this approach fares best overall. These results show that active view selection enhances the robot's ability to learn and recognize objects in real-world environments

chitecture LeNet [78.49] consists of a neural network that uses a hierarchy of layers of trainable convolutions and spatial subsampling, as well as nonlinear 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.* [78.50] 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, the 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.

Recent research has mainly focused on two topics: the automatic extraction of optimal visual features for efficient recognition and categorization, and the extension of the frameworks for providing invariance against changes in viewing conditions (such as rotations in depth, scaling, translation, illumination, and occlusion, for example, *DiCarlo et al.* [78.52], *Rolls* [78.53] for discussions of invariance in neuromorphic architectures). In a recent paper [78.54], these two issues have been addressed in a face recognition task conducted on a difficult database of faces taken in uncontrolled environments. The selection of optimal features was done by evaluating a large set of potential visual feature combinations using GPU-accelerated algorithms. The issue of invariance was addressed by using a hierarchical, multilayer model in which each layer includes linear and nonlinear pooling operations that encode the input image. The combined system was benchmarked against other standard feature-extraction methods and a *flat*, nonhierarchical one-layer model. Both properties resulted in increased recognition performance on the database outperforming other benchmarked state-of-the-art methods. In addition, the system also showed increased robustness against viewing variations, which included pose, position, scale, and background clutter.

In summary, neuromorphic architectures have now reached a stage of maturity that can put them even ahead of sophisticated, state-of-the-art computer vision frameworks. The ability to learn and adapt the feature set to viewing conditions and the increased robustness to viewing conditions makes such architectures good candidates for building the visual learning and recognition system for a perceptual robot.

78.1.4 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 feedforward – 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 Chap. 77. Although there is evidence that humans solve some recognition tasks using very little feedback (see, e.g., *Thorpe et al.* [78.51]; *DiCarlo et al.* [78.52]), 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 Chap. 77.

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, nonrigid 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 [78.55–58]. These results ask for an extension of current object recognition frameworks with a temporal component in order to arrive at truly spatiotemporal object representations. Combining methods from computer vision, psychophysics, and machine learning, *Wallraven and Bülthoff* [78.59, 60], have developed a framework that fulfills this requirement and learns spatiotemporal, exemplar-based object representations from image sequences. 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 [78.61]. In sev-

eral 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 spatiotemporal 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 similar model using a neuromorphic architecture integrating the temporal dimension was proposed by *Kietzmann et al.* [78.62].

A third issue that – in our view – will be essential for designing and implementing efficient perceptual robots consists of the multisensory nature of our perceptual system (see also the discussion of embodied robots in Chap. 13). 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 [78.63], participants had to learn views of four simple, 3-D 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 actively grasping and touching objects, which thus could provide a *database* of object representations for visual recognition [78.64]. Using this basic perceptual principle as a motivation [78.60] have applied an extension of the keyframe framework in an online robotics scenario for efficient learning and recognition of multisensory 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 spatiotemporal, view-based representations using the keyframe framework. 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. The framework was also used as the basis for recent work in which a humanoid robot (the iCub) performed active in-hand object recognition, searching for the optimal view that allowed it to disambiguate the object currently held from other, previously seen objects [78.47]. Again, linking the exploratory actions (turning the hand) with the visual data resulted in a much faster and more reliable object recognition performance. Sample data comparing unplanned and planned recognition of difficult objects is shown in Fig. 78.2 ( VIDEO 569).

Such approaches pave the way for a view of recognition as an active, multisensory process in which rich, extensible object representations are formed and improved over the life-time of the robot.

78.2 Perceptual Mechanisms of Action Representation

The recognition of complex movements and actions is fundamental for many applications in robotics, such as imitation learning by observation. Interactive robots need to analyze their users' movements in order to respond in a natural way to their social and emotional behavior (Chap. 72). The following section reviews what is known about movement and action recognition in the brain and tries to highlight a few aspects that have

or might be successfully transferred to biologically inspired applications in robotics and computer vision.

78.2.1 Recognition of Complex Movements and Actions in Primate Cortex

The recognition of complex movements and actions is a fundamental problem for higher animals and specif-

ically for primates. While simple movement patterns are sufficient for eliciting stereotypical prey catching behavior in simple vertebrates ([78.65]; see also Chap. 77), higher animals exploit more complex movement patterns, e.g., for the recognition of conspecifics or predators, or for communication by facial movements, gestures, or body expressions. Human perception of body motion patterns is very efficient, even for extremely impoverished stimuli. This has been demonstrated in classical experiments by *Johansson* [78.66], 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 (Sect. 78.1.2), some studies have tried to uncover neural [78.67–70], and computational principles [78.71–73] of visual movement recognition. Some of these principles have been transferred to the construction of systems in computer vision and robotics.

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 Sect. 78.1.2. 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 Fig. 78.3. 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 [78.69], and similar structures are activated by these stimuli in the human brain. In addition, areas selective for human body shapes, such as the *extrastriate body part area* (EBA), likely to contribute to the recognition of actions [78.74], where information of form and motion features seems to be integrated on higher processing levels [78.73].

For the recognition of goal-directed actions, such as reaching or grasping, in addition cortical structures beyond the visual cortex, such as the parietal and premotor cortex, seem to play a critical role. The role of these

structures for action recognition has been analyzed in particular in the context of the study of the *mirror neuron system* [78.75]. Mirror neurons are sensorimotor neurons that combine visual tuning during action observation as well as selective motor tuning. Areas in parietal cortex, such as the anterior might be specifically relevant for the recognition of action-related objects and their relationship to moving effectors [78.36]. Research about the mirror neuron system has influenced the construction of a whole generation of biologically inspired robots (Chap. 77). The guiding hypothesis has been that the visual recognition and *understanding* of actions is accomplished by mapping of observed body movements onto motor representations that are relevant for the execution of the same type of action.

78.2.2 Biological Principles with Relevance for Computer Vision and Robotics

We discuss in the following two major principles that have been derived from the analysis of action recognition in biological systems that have been transferred to technical applications in computer vision and robotics.

A first principle that seems to be implemented in movement recognition in primate cortex is a hierarchical architecture of feature detectors, which accomplishes action recognition by the detection of temporal sequences of relevant motion and form features. Such detection does not necessarily require the reconstruction of the three-dimensional facial or body shape, nor an exact simulation of the dynamics of the underlying movements. Instead it can be accomplished by much simpler computational mechanisms. Like object recognition, the recognition of complex motion patterns is strongly orientation- and view-dependent. This property has been observed at the level of individual neurons in the STS, and for the activation of biological motion-selective areas in human cortex [78.69] as well as for action-selective neurons in higher areas such as the premotor cortex [78.76]. View- and orientation dependence seem compatible with an encoding of visually perceived movements in terms of potentially learned example views, or *keyframes* (snap shots), and of instantaneous optic flow patterns that are characteristic for actions [78.67, 73]. While there might be some innate preferences for specific features [78.77], psychophysical and fMRI (functional magnetic resonance imaging) experiments suggest an important role of learning in visual movement recognition [78.78–80]. For example, subjects can learn easily to recognize individually – specific body and facial movements [78.81, 82]. Learning-based theoretical models, exploiting similar principles as neural object recognition models, account for a variety of experimental data on action

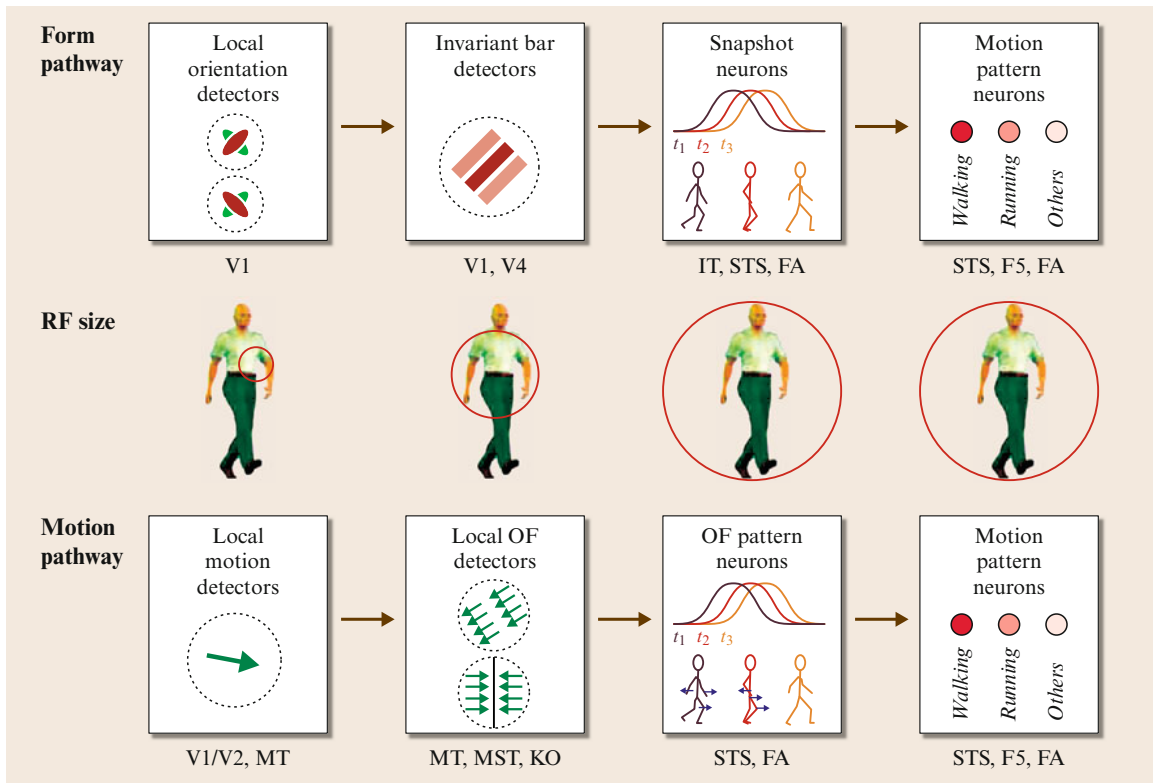


Fig. 78.3 Example-based neural model for the visual recognition of body movements that integrates the processing of form and motion features in the ventral and dorsal visual pathways (after *Giess and Poggio* [78.73])

recognition in biological systems [78.73, 83, 84], also supporting a central role of learning.

As an example of such a learning-based architecture, Fig. 78.3 illustrates a hierarchical model for the recognition of complex body movements [78.73]. It 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 Sect. 78.1.2. 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 (Chap. 68). A strong activity in the network emerges only when the stimulus sequence matches these predictions.

Similar hierarchical neural architectures inspired by the visual cortex have been used in the context of mirror-neuron robot systems [78.85, 86]. In addition, recent work in computer vision shows that such biologically inspired architectures can reach very high performance levels, comparable to state-of-the-art algorithms in computer vision [78.87–89].

A second principle of movement recognition, which has been discussed extensively as basis for the recognition of imitable actions, and as explanation of the function of the mirror neuron system [78.75] is the idea that action observation is based on an internal simulation of the observed motor behavior. A variety of computational models for action recognition by internal simulation have been proposed in the neuroscience literature, e.g., exploiting feedforward controllers [78.90], coupled forward and backward models [78.91], hierarchical Bayesian predictive models [78.92], or a free energy minimization framework [78.93]. (A further more extensive discussion about theoretical models for the mirror neuron system with relevance for robotics

can be found in Chaps. 68 and 77.) A main difficulty of the recognition of actions by internal simulation of associated motor behaviors is the accurate estimation of relevant geometrical quantities from image data, especially when no special depth sensors or even online motion capture are available. Many of the underlying motor control models are formulated in joint angle space, and the robust recognition of joint angles from monocular videos is known to be a difficult computer vision problem, which so far is solvable only for highly restricted classes of movements with strong learned priors, and at considerable computational cost [78.94, 95]. This raises the question about simpler computational approaches for the recognition of goal-directed actions, which explain biological data and might be interesting for technical applications.

A recently developed model for the visual recognition of goal-directed hand actions in cortex that follows these lines [78.96] is illustrated in Fig. 78.4. The underlying architecture is an extension of the form pathway of the model shown in Fig. 78.3, by the addition of neural circuits that process the spatial and temporal relationship between the observed effector (in this case the hand) and the recognized goal object (e.g., a grasped object). The model works, exploiting a purely exemplar-based approach (Sect. 78.1.1), without explicit reconstruction of the three-dimensional structure of the object or the effector. The model comprises three modules: The first module (A, in Fig. 78.4) recognizes shapes of the goal object and of the effector,

implementing a shape recognition hierarchy similar to standard object recognition models as the ones described in Sect. 78.1.3. The analysis of the temporal deformation of the hand is based on the recognition of sequences of key shapes, like in the form pathway of the model in Fig. 78.3. Opposed to standard object recognition models, however, the highest level of this shape recognition hierarchy is not completely position-invariant. Rather, it retains coarse position selectivity by implementing multiple replica of the same shape detectors that are selective for different image positions. This makes it possible to further analyze the spatiotemporal relationship of the recognized goal objects and effector. This analysis is realized in the second module (B) whose core is formed by two-dimensional *relative position maps*. These are neural activity maps that represent the effector position as activity peak in a two-dimensional coordinate system that is centered on the object position in the image. These maps are computed by a gain fields [78.97] that multiply the output activities of shape selective neurons with selectivity for object and effector. The activity distribution in these neural map is analyzed by *affordance neurons* that are activated only when hand and object shape match and are in a spatial relationship that is suitable for a successful grip. By appropriate pooling of the responses of motion energy detectors that receive input from the relative position map *relative motion neurons* can be constructed, whose activity characterizes the relative motion between the effector and object

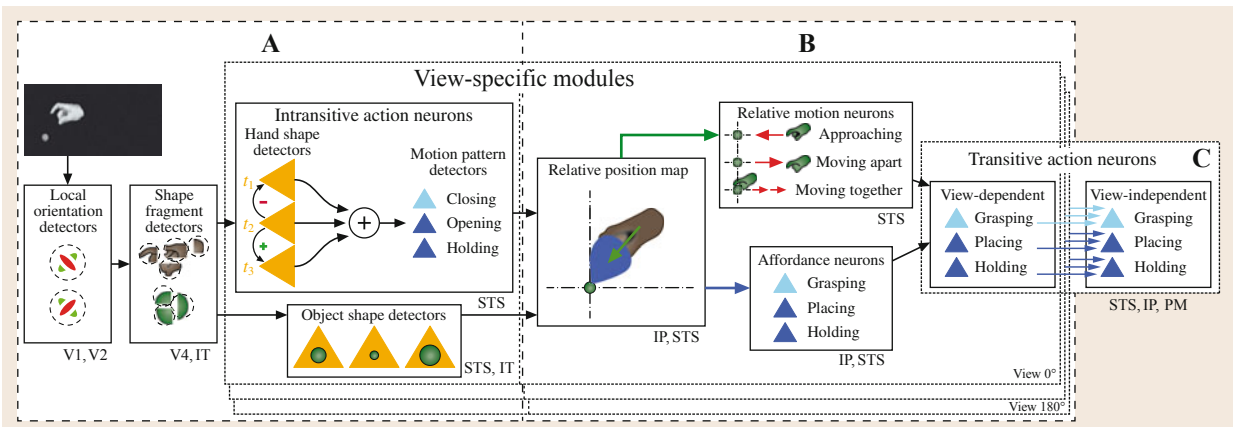


Fig. 78.4a–c Physiologically inspired model for the recognition of goal-directed hand actions. The shape-recognition (a) recognizes the shape of goal objects and the shapes of individual hand postures, retaining some coarse position information. The module (b) associates the information of hand and object by computing maps that represent the relative positions of hand and object in image coordinates. From these maps the spatial matching hand and object and their relative motion can be computed. The highest level module (c) contains model neurons that are selective for different types of goal-directed actions. Up to this level the model recognizes actions in a view-dependent manner, and only at the highest level (view-independent transitive actions neurons) the model accomplishes view independence by pooling the outputs from view-specific modules (courtesy of *Fleischer et al.* [78.96])

(e.g., the hand approaching the object). The outputs of affordance and the relative motion neurons are integrated in the third module (C), which contains only neurons that are selective for visually observed goal-directed actions. This integration of information is first accomplished in a purely view-specific manner. View independence is not established until the very last hierarchy level of the model that contains view-independent action-selective neurons. The idea of establishing view-dependence only very late in the cortical hierarchy seems counter-intuitive, and is at odds with several established computational models for action recognition. However, this dominance of exemplar-based representations until very high levels of the cortical processing hierarchy has been observed in electrophysiological experiments studying mirror neurons in premotor cor-

tex [78.76]. In this structure, which is traditionally associated with motor planning, the majority of mirror neurons is view-dependent and only a minority is view-independent. The discussed model can recognize hand actions from gray level videos. It could be augmented by integration of disparity or depth features, and by appropriate attentional control mechanisms that would make it more robust to cluttered scenes with multiple relevant objects. Whether similar architectures have advantages for the robust visual recognition of goal-directed actions in technical systems remains to be shown. Very recent work shows that such hierarchical deep architectures, which consist of learned feature detectors, outperform classical technical solutions on actual computer vision benchmarks for action detection [78.98].

78.3 Perceptual Validation of Robotics

Successful human–robot interaction is perhaps easiest when the robot offers interaction channels that are compatible to that of human–human interaction [78.99]. The most important interaction channels in this case are verbal and nonverbal communication with the face. Importantly, in human–human interaction, nonverbal communication using facial expressions, for example, constitutes up to 30% of the communicative content. Facial expressions are not only used to convey someone’s mood and emotion [78.100], but are also used in communicative contexts to signal understanding (a nod of the head), to modify what is being said (a raise of the eye-brows), and to control the conversational flow (a look of confusion may signal to the speaker to repeat what has been said). Hence, many humanoids have incorporated more or less sophisticated heads capable of producing human-like facial expressions and movements. Traditionally, this has been achieved using mechatronic implementations in which actuators drive facial features directly (e.g., as in the MDS robot by Lee and Breazeal [78.101]), or – in more complex implementations of android robots – mimic human muscle movements that are then used to deform artificial skin [78.102–104]. Other systems have used LEDs for displaying simple, changeable facial features (e.g., as in the iCub platform [78.105].

78.3.1 Realistic Faces for Robots

With such a great variety of robot systems also comes the need for an investigation of their perceptual evaluation and their interaction capabilities ([78.106] for such a study in the context of facial animations in

computer graphics). Indeed, one particular problem of android, human-like systems is that they easily could suffer from the uncanny valley effect as the actuators and/or the control framework cannot easily reproduce the smoothness of human facial expressions. A study of morphed images between a nonhuman robot face and a highly realistic android robot head, for example, clearly showed evidence of the uncanny valley effect [78.107] – a similar study for moving robot faces yielded more mixed results, but still showed that the most realistic robotic faces were clearly perceived as different from that of a human talking [78.108]. One solution for this is to change the robot’s appearance such that it stays away from close human likeness; however, conveying the full breadth of human communicative signals with different facial features, or a different facial topology may also be problematic. A different solution consists of avoiding a mechanical solution and instead resorting to facial animation from computer graphics. One example of such a system was presented in *De-launay et al.* [78.109] in which facial animations are projected onto a rigid face mask. Since facial animation techniques are in many ways much more advanced, such a system allows for a more realistic and flexible interaction in human–robot interaction. Subtle cues such as eye-gaze, wrinkles, and other nonrigid facial deformations could be displayed via projection of an advanced facial animation engine. Initial perceptual experiments with such systems [78.109, 110] have yielded promising results. However, more studies need to be done to assess the properties of human–robot interactions in these and other implementations of facial displays.

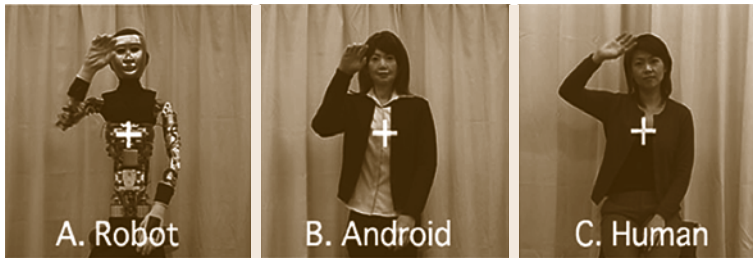


Fig. 78.5 Stimuli used in experiment investigating the fMRI correlates of the observation of human and robot movements. The neural responses to the movements by a real human are compared with the ones induced by a human-looking robot (Android), and by a nonhumanlike robot (Robot) (after Saygin et al. [78.111])

78.3.2 Perceptual and Neural Processing of Body Movements of Robots

Research in humanoid robotics finally aims at optimizing the perceived naturalness or *human-likeness* of generated robot movements, since this in the long run will increase the acceptance of humanoid robots in social contexts. However, the present humanoid platforms have typically substantial constraints that still prevent the realization of complex really human-like movements. This is even more the case for the realization of behaviors on bipedal robots, due to the difficult problem to maintain dynamic balance (Chap. 67). Therefore, most body movements realized by present humanoid robots still differ in many aspects from human movements. This makes the quantification of the degree of realism of such movements presently a less pressing topic than the field of computer graphics, where psychophysical studies for the validation of the realism and quality of computer animation methods are meanwhile a standard [78.112–114]. However, research in psychophysics and neuroscience has started to investigate the differences between the perceptual processing of human and robot movements, and interesting results have been obtained that localize cortical subsystems that might be essential to distinguish human and nonhuman robot movements. A typical question in these studies has been which critical properties determine whether visual stimuli produce *motor resonance*, or an activation of action-selective neural structures. The results of such studies have not been completely consistent, since some studies found decreased activation of action-selective networks for robot movements [78.115–118] while others found no such differences [78.119]. Primarily visual processing areas responded sometimes more for robot movements than for normal human [78.111, 117].

The problem of such studies is that many factors might influence the perception and neural signals in action-selective areas, such as form, kinematics, and optic flow patterns. Typically, it is very difficult to control these parameters separately for real robots. In addition, the learning experience of observers with the specific robot might play an important role [78.120]. A recent study by Saygin et al. [78.111] tried to separate at least the influences of the robot appearance (shape) and the motion kinematics by comparing the fMRI signals (using an adaptation paradigm) driven by three different stimuli (Fig. 78.5): a real person (that served as model for the building of the robot), the human-like looking robot (*android*), and the robot without skin and surface parts that made it look human-like (resulting in very similar motion as the full robot). In visual areas (e.g., the extrastriate body area) the human and the human-like robot stimulus result in very similar activity. This is not true for the parietal cortex, which is part of the *mirror neuron network*. This region shows large differences between the human-like robot and the other two conditions, potentially reflecting an increase of neural processing resources that are required to cope with the contradiction between the form and the kinematic information that is presented by this stimulus. Opposed to this, the not human-like robot makes it expected that the motion is also not human-like, potentially causing no such conflict. Future studies of similar type, controlling for the different information channels of action processing (Sect. 78.2.2) as well as for the predictability of such stimuli dependent on previous learning, potentially combined with quantitative neural modeling, will be required to really understand how different factors are integrated in the neural processing of robot movements, and how this causes different levels of perceived *human-likeness*.

78.4 Conclusion and Further Reading

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 these brain functions are likely realized by example-based representations. We have discussed neural implementations of such representations which partially have been tested successfully in the context of technical applications, and which are strongly inspired by the real cortical neural architecture. In addition, we have presented some new computational principles that seem to emerge from recent experimental results on the representation of goal-directed actions. Finally, we have discussed work that tries to use psychophysical and neuroscience methods for the validation of the appearance and the movements of human-like robots, and for the investigation of underlying neural mechanisms.

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 (Chaps. 36–38, 47). 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) [78.121]. However, it is unclear whether the information about objects is only represented in terms of 2-D example views. Instead, it seems likely that also some form of 3-D 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 [78.122]. A further discussion about biologically inspired models for the extraction of action-relevant geometrical information in the context of grasping and manipulation is given in Chaps. 32 and 77.

The perceptual validation of the human-likeness and affective impact of humanoid robots likely will become increasingly important along with the further development of the technology that will increase the level of similarity between humanoid robots and humans. Likewise, it seems increasingly important to use quantitative methods from perception science to investigate the quality of the emotional and social interaction between robots and humans. We expect this to be a field where psychology can really contribute quantitative methods to engineering, reaching a level that goes beyond a qualitative and subjective comparison of *demos* which partially is still the standard in the field of humanoid robotics.

Video-References

 VIDEO 569 Active in-hand object recognition available from <http://handbookofrobotics.org/view-chapter/78/videodetails/569>

References

- | | |
|--|---|
| <p>78.1 D. Marr: <i>Vision</i> (Freeman, San Francisco 1982)</p> <p>78.2 J.A.S. Kelso: <i>Dynamic Patterns: The Self-Organization of Brain and Behaviour</i> (MIT, Cambridge 1995)</p> <p>78.3 G. Schöner, M. Dose, C. Engels: Dynamics of behavior: Theory and applications for autonomous robot architectures, <i>Robotics Auton. Syst.</i> 16, 213–245 (1997)</p> <p>78.4 J. Tani, M. Ito: Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment, <i>IEEE Trans. Syst. Man Cybern. A</i> 33(4), 481–488 (2003)</p> <p>78.5 W.H. Warren: The dynamics of perception and action, <i>Psychol. Rev.</i> 113, 358–389 (2006)</p> <p>78.6 M. Mori: The uncanny valley, <i>Energy</i> 7(4), 33–35 (1970), in Japanese</p> <p>78.7 D.W. Cunningham, C. Wallraven: <i>Experimental Design: From User Studies to Psychophysics</i> (CRC, Boca Raton 2011)</p> | <p>78.8 A. Field, G. Hole: <i>How to Design and Report Experiments</i> (Sage, London 2011)</p> <p>78.9 D. Marr, H. Nishihara: Representation and recognition of the spatial organization of three-dimensional shapes, <i>Proc. R. Soc. B</i> 200, 269–294 (1978)</p> <p>78.10 I. Biederman: Recognition-by-components: A theory of human image understanding, <i>Psychol. Rev.</i> 94, 115–147 (1987)</p> <p>78.11 D. Lowe: <i>Perceptual Organization and Visual Recognition</i> (Kluwer, Boston 1985)</p> <p>78.12 S. Ullman: <i>High-Level Vision. Object Recognition and Visual Cognition</i> (MIT, Cambridge 1996)</p> <p>78.13 M.A. Kurbat: Structural description theories: Is RBC/JIM a general-purpose theory of human entry-level object recognition?, <i>Perception</i> 23, 1339–1368 (1994)</p> <p>78.14 S. Edelman: <i>Representation and Recognition in Vision</i> (MIT, Cambridge 1999)</p> |
|--|---|

- 78.15 M. Graf, W. Schneider: Structural descriptions in HIT – A problematic commitment, *Behav. Brain Sci.* **24**, 483–484 (2001)
- 78.16 H.H. Bülthoff, S. Edelman: Psychophysical support for a two-dimensional view interpolation theory of object recognition, *Proc. Natl. Acad. Sci. USA* **89**, 60–64 (1992)
- 78.17 M.J. Tarr, S. Pinker: Mental orientation and orientation-dependence in shape recognition, *Cogn. Psychol.* **21**, 233–282 (1989)
- 78.18 W.G. Hayward, M.J. Tarr: Testing conditions for viewpoint invariance in object recognition, *J. Exp. Psychol.* **23**, 1511–1521 (1997)
- 78.19 S.E. Palmer, E. Rosch, P. Chase: Canonical perspective and the perception of objects. In: *Attention and Performance IX*, ed. by J. Long, A. Baddeley (Erlbaum, Hillsdale 1981) pp. 135–151
- 78.20 H. Hill, P.G. Schyns, S. Akamatsu: Information and viewpoint dependence in face recognition, *Cognition* **62**, 201–222 (1997)
- 78.21 C. Wallraven, A. Schwaninger, S. Schuhmacher, H.H. Bülthoff: View-based recognition of faces in man and machine: Re-visiting inter-extra-ortho, *Lect. Notes Comput. Sci.* **2525**, 651–660 (2002)
- 78.22 M.J. Tarr: Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects, *Psychon. Bull. Rev.* **2**, 55–82 (1995)
- 78.23 R. Lawson, G.W. Humphreys: View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects, *Percept. Psychophys.* **60**, 1052–1066 (1998)
- 78.24 E. Ashbridge, D.I. Perrett: Generalizing across object orientation and size. In: *Perceptual Constancy. Why Things Look as They Do*, ed. by V. Walsh, J. Kulikowski (Cambridge Univ. Press, Cambridge 1998) pp. 192–209
- 78.25 M. Dill, S. Edelman: Imperfect invariance to object translation in the discrimination of complex shapes, *Perception* **30**, 707–724 (2001)
- 78.26 K.R. Cave, S. Pinker, L. Giorgi, C.E. Thomas, L.M. Heller, J.M. Wolfe, H. Lin: The representation of location in visual images, *Cogn. Psychol.* **26**, 1–32 (1994)
- 78.27 S. Ullman: Aligning pictorial descriptions: An approach to object recognition, *Cognition* **32**, 193–254 (1989)
- 78.28 S. Ullman, R. Basri: Recognition by linear combinations of models, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 992–1006 (1991)
- 78.29 T. Poggio, S. Edelman: A network that learns to recognize three-dimensional objects, *Nature* **343**, 263–266 (1990)
- 78.30 D. Perrett, W.M. Oram: Visual recognition based on temporal cortex cells: Viewer-centred processing of pattern configurations, *Z. Naturforsch. C* **53**, 518–541 (1998)
- 78.31 M. Riesenhuber, T. Poggio: Hierarchical models of object recognition in cortex, *Nat. Neurosci.* **2**, 1019–1025 (1999)
- 78.32 E.T. Rolls, T. Milward: A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures, *Neural Comput.* **2**(11), 2547–2572 (2000)
- 78.33 G. Wallis, H.H. Bülthoff: Learning to recognize objects, *Trends Cogn. Sci.* **3**, 22–31 (1999)
- 78.34 D. Perrett, W.M. Oram, E. Ashbridge: Evidence accumulation in cell populations responsive to faces: An account of generalization of recognition without mental transformations, *Cognition* **67**, 111–145 (1998)
- 78.35 D.H. Foster, S.J. Gilson: Recognizing novel three-dimensional objects by summing signals from parts and views, *Proc. R. Soc. B* **269**, 1939–1947 (2002)
- 78.36 H. Sakata: The role of the parietal cortex in grasping, *Adv. Neurol.* **93**, 121–139 (2003)
- 78.37 D.H. Hubel, T.N. Wiesel: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol. (Lond.)* **160**, 106–154 (1962)
- 78.38 G. Wang, M. Tanifuji, K. Tanaka: Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging, *Neurosci. Res.* **32**, 33–46 (1998)
- 78.39 K. Tanaka: Representation of visual feature objects in the inferotemporal cortex, *Neural Netw.* **9**(8), 1459–1475 (1996)
- 78.40 K. Grill-Spector, R. Malach: The human visual cortex, *Annu. Rev. Neurosci.* **27**, 649–677 (2004)
- 78.41 N.K. Logothetis, J. Pauls, H.H. Bülthoff, T. Poggio: View-dependent object recognition by monkeys, *Curr. Biol.* **4**, 401–414 (1994)
- 78.42 L. Roberts: Machine perception of three-dimensional solids. In: *Optical and Electro-Optical Information Processing*, ed. by J.T. Tippet (MIT, Cambridge 1965) pp. 159–197
- 78.43 M. Swain, D. Ballard: Color indexing, *Int. J. Comput. Vis.* **7**, 11–32 (1991)
- 78.44 C. Schmid, R. Mohr: Local greyvalue invariants for image retrieval, *IEEE Trans. Pattern Mach. Intell.* **19**, 530–535 (1997)
- 78.45 D. Lowe: Distinctive image features from scale invariant keypoints, *Int. J. Comput. Vis.* **60**(2), 90–110 (2004)
- 78.46 M. Kirby, L. Sirovich: Applications of the Karhunen–Loeve procedure for the characterization of human faces, *IEEE Trans. Pattern Mach. Intell.* **12**, 103–108 (1990)
- 78.47 B. Browatzki, V. Tikhonoff, G. Metta, H.H. Bülthoff, C. Wallraven: Active in-hand object recognition on a humanoid robot, *IEEE Trans. Robotics* **30**(5), 1260–1269 (2014)
- 78.48 A. Delorme, S. Thorpe: SpikeNET: An event-driven simulation package for modeling large networks of spiking neurons, *Netw. Comput. Neural Syst.* **14**, 613–627 (2003)
- 78.49 Y. LeCun, F. Huang, L. Bottou: Learning methods for generic object recognition with invariance to pose and lighting, *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.* (2004)
- 78.50 T. Serre, L. Wolf, T. Poggio: Object recognition with features inspired by visual cortex, *Proc. 2005 IEEE*

- Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2005)
- 78.51 S. Thorpe, D. Fize, C. Marlot: Speed of processing in the human visual system, *Nature* **381**(6582), 520–522 (1996)
- 78.52 J.J. DiCarlo, D. Zoccolan, N.C. Rust: How does the brain solve visual object recognition?, *Neuron* **73**(3), 415–434 (2012)
- 78.53 E.T. Rolls: Invariant visual object and face recognition: Neural and computational bases, and a model, *VisNet. Front. Comput. Neurosci.* **6**(35), (2012)
- 78.54 N. Pinto, D. Cox: High-throughput-derived biologically-inspired features for unconstrained face recognition, *Image Vis. Comput.* **30**(3), 159–168 (2012)
- 78.55 G. Wallis, H.H. Bülthoff: Effects of temporal association on recognition memory, *Proc. Natl. Acad. Sci. USA* **98**, 4800–4804 (2001)
- 78.56 J.V. Stone: Object recognition using spatio-temporal signatures, *Vis. Res.* **38**(7), 947–951 (1998)
- 78.57 J.V. Stone: Object recognition: View-specificity and motion-specificity, *Vis. Res.* **39**(24), 4032–4044 (1999)
- 78.58 Q.C. Vuong, M.J. Tarr: Rotation direction affects object recognition, *Vis. Res.* **44**(14), 1717–1730 (2004)
- 78.59 C. Wallraven, H.H. Bülthoff: Automatic acquisition of exemplar-based representations for recognition from image sequences, *CVPR 2001 – Workshop Models vs. Ex.* (2001)
- 78.60 C. Wallraven, H.H. Bülthoff: Object recognition in humans and machines. In: *Object Recognition, Attention and Action*, ed. by N. Osaka, I. Rentschler, I. Biederman (Springer, Tokyo 2007) pp. 89–104
- 78.61 C. Wallraven, B. Caputo, A.B.A. Graf: Recognition with local features: The kernel recipe, *Proc. Int. Conf. Comput. Vis.*, Vol. 2 (2003) pp. 257–264
- 78.62 T.C. Kietzmann, S. Lange, M. Riedmiller: Computational object recognition: A biologically motivated approach, *Biol. Cybern.* **100**, 59–79 (2009)
- 78.63 F.N. Newell, M.O. Ernst, B.S. Tjan, H.H. Bülthoff: Viewpoint dependence in visual and haptic object recognition, *Psychol. Sci.* **12**, 37–42 (2001)
- 78.64 H. Lee, C. Wallraven: Exploiting object constancy: Effects of active exploration and shape morphing on similarity judgments of novel objects, *Exp. Brain Res.* **225**(2), 277–289 (2012)
- 78.65 J.-P. Ewert: Neural mechanisms of prey-catching and avoidance behavior in the toad *Bufo bufo* L, *Brain Behav. Evol.* **3**, 36–56 (1970)
- 78.66 G. Johansson: Visual perception of biological motion and a model for its analysis, *Percept. Psychophys.* **14**, 201–211 (1973)
- 78.67 K. Verfaillie: Perceiving human locomotion: Priming effects in direction discrimination, *Brain Cogn.* **44**, 192–213 (2000)
- 78.68 A.J. O’Toole, D.A. Roark, H.H. Abdi: Recognizing moving faces: A psychological and neural synthesis, *Trends Cogn. Sci.* **6**, 261–266 (2002)
- 78.69 D. Perrett, A. Puce: Electrophysiology and brain imaging of biological motion, *Philos. Trans. R. Soc. B* **358**, 435–445 (2003)
- 78.70 R. Blake, M. Shiffrar: Perception of human motion, *Annu. Rev. Psychol.* **58**, 47–73 (2007)
- 78.71 D.D. Hoffman, B.E. Flinchbaugh: The interpretation of biological motion, *Biol. Cybern.* **42**, 195–204 (1982)
- 78.72 J.A. Webb, J.K. Aggarwal: Structure from motion of rigid and jointed objects, *Artif. Intell.* **19**, 107–130 (1982)
- 78.73 M.A. Giese, T.T. Poggio: Neural mechanisms for the recognition of biological movements, *Nat. Rev. Neurosci.* **4**, 179–192 (2003)
- 78.74 J. Jastorff, G.A. Orban: Human functional magnetic resonance imaging reveals separation and integration of shape and motion cues in biological motion processing, *J. Neurosci.* **29**(22), 7315–7329 (2009)
- 78.75 G. Rizzolatti, L. Craighero: The mirror-neuron system, *Annu. Rev. Neurosci.* **27**, 169–192 (2004)
- 78.76 V. Caggiano, L. Fogassi, G. Rizzolatti, J.K. Pomper, P. Thier, M.A. Giese, A. Casile: View-based encoding of actions in mirror neurons of area f5 in macaque premotor cortex, *Curr. Biol.* **21**(2), 144–148 (2011)
- 78.77 F. Simion, E. Di Giorgio, I. Leo, L. Bardi: The processing of social stimuli in early infancy: From faces to biological motion perception, *Prog. Brain Res.* **189**, 173–193 (2011)
- 78.78 E.D. Grossman, R. Blake, C.Y. Kim: Learning to see biological motion: Brain activity parallels behavior, *J. Cogn. Neurosci.* **16**, 1669–1679 (2004)
- 78.79 J. Jastorff, Z. Kourtzi, M.A. Giese: Learning to discriminate complex movements: Biological versus artificial trajectories, *J. Vis.* **6**, 791–804 (2006)
- 78.80 J. Jastorff, Z. Kourtzi, M.A. Giese: Visual learning shapes the processing of complex movement stimuli in the human brain, *J. Neurosci.* **29**(44), 14026–14038 (2009)
- 78.81 H. Hill, F.E. Pollick: Exaggerating temporal differences enhances recognition of individuals from point light displays, *Psychol. Sci.* **11**, 223–228 (2000)
- 78.82 B. Knappmeyer, I.M. Thornton, H.H. Bülthoff: The use of facial motion and facial form during the processing of identity, *Vis. Res.* **43**, 1921–1936 (2003)
- 78.83 J. Lee, W. Wong: A stochastic model of coherent motion detection, *Biol. Cybern.* **91**, 306–314 (2004)
- 78.84 J. Lange, M. Lappe: A model of biological motion perception from configurational form cues, *J. Neurosci.* **26**(11), 2894–2906 (2006)
- 78.85 G. Tessitore, F. Donnarumma, R. Prevete: An action-tuned neural network architecture for hand pose estimation, *Proc. Int. Conf. Fuzzy Comput. Int. Conf. Neural Comput. Valencia* (2010) pp. 358–363
- 78.86 G. Metta, G. Sandini, L. Natale, L. Craighero, L. Fadiga: Understanding mirror neurons – A bio-robotic approach, *Interact. Stud.* **7**, 197–232 (2006)

- 78.87 H. Jhuang, T. Serre, L. Wolf, T. Poggio: A biologically inspired system for action recognition, *IEEE Int. Conf. Comput. Vis. (ICCV)* (2007) pp. 1–18
- 78.88 M.J. Escobar, G.S. Masson, T. Vieville, P. Kornprobst: Action recognition using a bio-inspired feedforward spiking network, *Int. J. Comput. Vis.* **82**(3), 284–301 (2009)
- 78.89 H. Jhuang, E. Garrote, J. Mutch, T. Poggio, A. Steele, T. Serre: Automated home-cage behavioral phenotyping of mice, *Nat. Commun.* **1**(86), 1–9 (2010)
- 78.90 D.M. Wolpert, K. Doya, M. Kawato: A unifying computational framework for motor control and social interaction, *Philos. Trans. R. Soc. B* **358**, 593–602 (2003)
- 78.91 Y. Demiris, M. Johnson: Distributed, predictive perception of actions: A biologically inspired robotics architecture for imitation and learning, *Connect. Sci.* **15**(4), 231–243 (2003)
- 78.92 J.M. Kilner, K.J. Friston, C.D. Frith: The mirror-neuron system: A Bayesian perspective, *Neuroreport* **18**, 619–623 (2007)
- 78.93 K. Friston, J. Mattout, J. Kilner: Action understanding and active inference, *Biol. Cybern.* **104**(1/2), 137–160 (2011)
- 78.94 R. Li, T.P. Tian, S. Sclaroff, M.H. Yang: 3D human motion tracking with a coordinated mixture of factor analyzers, *Int. J. Comput. Vis.* **87**, 170–190 (2010)
- 78.95 D.R. Weinland, R. Ronfard, E. Boyer: A survey of vision-based methods for action representation. Segmentation and recognition, *Comput. Vis. Image Underst.* **115**(2), 224–241 (2011)
- 78.96 F. Fleischer, V. Caggiano, P. Thier, M.A. Giese: Physiologically inspired model for the visual recognition of transitive hand actions, *J. Neurosci.* **33**, 6563–6580 (2013)
- 78.97 E. Salinas, L.F. Abbott: Transfer of coded information from sensory to motor networks, *J. Neurosci.* **15**, 6461–6474 (1995)
- 78.98 A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei: Large-scale video classification with convolutional neural networks, *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, New York (2014) pp. 1725–1732
- 78.99 C. Breazeal: *Designing Sociable Robots* (MIT Press, Cambridge 2002)
- 78.100 M. Nusseck, D.W. Cunningham, C. Wallraven, H.H. Bülthoff: The contribution of different facial regions to the recognition of conversational expressions, *J. Vis.* **8**(8), 1 (2008)
- 78.101 J.K. Lee, C. Breazeal: Human social response toward humanoid robot's head and facial features, *Proc. CHI 2010* (2010) pp. 4237–4242
- 78.102 D. Hanson: Exploring the aesthetic range for humanoid robots, *CogSci-2006 Workshop: Toward Soc. Mech. Android Sci.* (2006)
- 78.103 H. Ishiguro: Understanding humans by building androids, *Proc. SIGDIAL Conf.* (2010)
- 78.104 P. Jaeckel, N. Campbell, C. Melhuish: Facial behaviour mapping - From video footage to a robot head, *Robotics Auton. Syst.* **56**(12), 1042–1049 (2008)
- 78.105 G. Metta, G. Sandini, D. Vernon, L. Natale, F. Nori: The iCub humanoid robot: An open platform for research in embodied cognition, *Proc. 8th Workshop Perform. Metr. Intell. Syst.* (2008) pp. 50–56
- 78.106 C. Wallraven, M. Breidt, D.W. Cunningham, H.H. Bülthoff: Evaluating the perceptual realism of animated facial expressions, *ACM Trans. Appl. Percept.* **4**(4), 1–20 (2008)
- 78.107 K.F. MacDorman: Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley, *ICCS/CogSci-2006 Symp. Toward Soc. Mech. Android Sci.* (2006) pp. 26–29
- 78.108 C. Ho, K.F. MacDorman, Z.A.D. Pramono: Human emotion and the uncanny valley: A GLM, MDS, and isomap analysis of robot video ratings, *Proc. HRI 2008* (2008) pp. 169–176
- 78.109 F. Delaunay, J. de Greeff, T. Belpaeme: Towards retro-projected robot faces: An alternative to mechatronic and android faces, *Proc. 18th IEEE Int. Symp. Robot Human Interact. Commun. RO-MAN* (2009) pp. 306–311
- 78.110 T. Kuratate, M. Riley, B. Pierce, G. Cheng: Gender identification bias induced with texture images on a life size retro-projected face screen, *Proc. 21st IEEE Int. Symp. Robot Human Interact. Commun. RO-MAN 2012* (2012) pp. 43–48
- 78.111 A.P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith: The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions, *Soc. Cogn. Affect. Neurosci.* **7**(4), 413–422 (2012)
- 78.112 P.S.A. Reitsma, N.S. Pollard: Perceptual metrics for character animation: Sensitivity to errors in ballistic motion, *ACM SIGGRAPH 2003 Papers (SIGGRAPH '03)* (ACM, New York 2003) pp. 537–542
- 78.113 T. Ezzat, G. Geiger, T. Poggio: Trainable videorealistic speech animation, *Proc. 29th Annu. Conf. Comput. Gr. Interact. Techn. (SIGGRAPH '02)* (ACM, New York 2002) pp. 388–398
- 78.114 J. Wang, B. Bodenheimer: Synthesis and evaluation of linear motion transitions, *ACM Trans. Graph.* **27**(1), Article 1 (2008)
- 78.115 M. Candidi, C. Urgesi, S. Ionta, S.M. Aglioti: Virtual lesion of ventral premotor cortex impairs visual perception of biomechanically possible but not impossible actions, *Soc. Neurosci.* **3**(3/4), 388–400 (2008)
- 78.116 T. Chaminade, J. Hodgins, M. Kawato: Anthropomorphism influences perception of computer-animated characters' actions, *Soc. Cogn. Affect. Neurosci.* **2**(3), 206–216 (2007)
- 78.117 T. Chaminade, M. Zecca, S.J. Blakemore, A. Takanishi, C.D. Frith, S. Micera, P. Dario, G. Rizzolatti, V. Gallese, M.A. Umiltà: Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures, *PLoS ONE* **5**(7), e11577 (2010)
- 78.118 Y.F. Tai, C. Scherfler, D.J. Brooks, N. Sawamoto, U. Castiello: The human premotor cortex is 'mir-

- ror' only for biological actions, *Curr. Biol.* **14**, 117–120 (2004)
- 78.119 L.M. Oberman, J.P. McCleery, V.S. Ramachandran, J.A. Pineda: EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots, *Neurocomput.* **70**, 2194–2203 (2007)
- 78.120 C. Press, H. Gillmeister, C. Heyes: Sensorimotor experience enhances automatic imitation of robotic action, *Proc. Biol. Sci.* **274**(1625), 2509–2514 (2007)
- 78.121 C.L. Colby: Action-oriented spatial reference frames in cortex, *Neuron* **20**, 15–24 (1998)
- 78.122 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 (2005)