13 Elements for a Neural Theory of the Processing of Dynamic Faces

Thomas Serre and Martin A. Giese

Face recognition has been a central topic in computer vision for at least two decades and progress in recent years has been significant. Automated face recognition systems are now widespread in applications ranging from surveillance to personal computers. In contrast, only a handful of neurobiologically plausible computational models have been proposed to try to explain the processing of faces in the primate cortex (e.g., Giese & Leopold, 2005; Jiang et al., 2006), and no such model has been applied specifically in the context of dynamic faces.

There is a need for an integrated computational theory of dynamic face processing that could integrate and summarize evidence obtained with different experimental methods, from single-cell physiology to fMRI, MEG, and ultimately behavior and psychophysics. At the same time, physiologically plausible models capable of processing real video sequences constitute a plausibility proof for the computational feasibility of different hypothetical neural mechanisms.

In this review chapter we will first discuss computer vision models for the processing of dynamic faces; these do not necessarily try to reproduce biological data but may suggest relevant computational principles. We then provide an overview of computational neuroscience models for the processing of static faces and dynamic body stimuli. We further highlight specific elements from our own work that are likely to be relevant for the processing of dynamic face stimuli. The last section discusses open problems and critical experiments from the viewpoint of neural computational approaches to the processing of dynamic faces.

Computational Models for the Processing of Faces and Bodies

The following section reviews work in computer vision as well as neural and psychological models for the recognition of static faces. In addition, we discuss neurobiological models for the processing of dynamic bodies. It seems likely that some of the computational principles proposed in the context of these models might also be relevant for the processing of dynamic faces.

Computer Vision Models for the Processing of Dynamic Faces

While a full review of the large body of literature on computer vision systems for the recognition and detection of static faces would far exceed the scope of this chapter (see Jain & Li, 2005; Kriegman, Yang, & Ahuja, 2002; Zhao, Chellappa, Rosenfeld, & Phillips, 2003 for relevant books and reviews), we discuss in the following pages a number of approaches for the recognition of facial expressions and dynamic facial stimuli. While these systems do not try to mimic the processing of information in the visual cortex, they do provide real-world evidence that critical information can be extracted from dynamic faces beyond the analysis of static frames.

Several computer vision systems have been developed for the recognition of facial expression based on the extraction of temporal cues from video sequences. First introduced by Suwa and colleagues (Suwa, Sugie, & Fujimora, 1978) in the 1970s and later popularized by Mase (Mase, 1991) in the 1990s, systems for the recognition of facial expressions have progressed tremendously in the past decades (see Fasel & Luettin, 2003; Pantic & Rothkrant, 2000; Tian, Kanade, & Cohn, 2005 for reviews).

Earlier approaches have typically relied on the computation of local optical flow from facial features (Black & Yacoob, 1995; Essa, Darrell, & Pentland, 1994; Mase, 1991; Otsuka & Ohya, 1996; Rosenblum, Yacoob, & Davis, 1994) and/or hidden Markov models (HMMs) to capture the underlying dynamics (Cohen, Sebe, Garg, Chen, & Huang, 2003; Otsuka & Ohya, 1996). More recent work (Pantic & Patras, 2006) has relied on the dynamics of individual facial points estimated using modern tracking algorithms. Another line of work (Yang, Liu, & Metaxas, 2009; Zhao et al., 2003) involves the extraction of image features shown to work well for the analysis of static faces (Ahonen, Hadid, & Pietikaine, 2006; Viola & Jones, 2001) across multiple frames.

Several behavioral studies (see for example chapters 2 and 4) have suggested that people might be able to extract idiosyncratic spatiotemporal signatures of a person's identity based on body and facial motion. An early approach that exploits the characteristic temporal signature of faces based on partially recurrent neural networks trained over sequences of facial images was first introduced by Gong, Psarrou, Katsoulis, and Palavouzis (1994). Initial experiments conducted by Luettin and colleagues suggested that spatiotemporal models (HMMs) trained on sequences of lip motion during speech could be useful for speaker recognition (Luettin, Thacker, & Beet, 1996). However, beyond this early experiment, the use of spatiotemporal cues for the identification of people in computer vision has remained relatively unexplored (Gong, McKenna, & Psarrou, 2000).

Overall, the success of computer vision systems for identifying people (i.e., face recognition) in video sequences, as opposed to static faces, has been more moderate (e.g., Edwards, Taylor, & Cootes, 1998; Lee, Ho, Yang, & Kriegman, 2003; Tistarelli, Bicego, & Grosso, 2009; Yamaguchi, Fukui, & Maeda, 1998). In fact, the idea of

exploiting video sequences for the recognition of faces was almost completely abandoned after it was concluded, from the face recognition vendor test (Phillips et al., 2002) (which offers an independent assessment of the performance of some of the leading academic and commercial face recognition systems), that the improvement from using video sequences over still images for face identification applications was minimal (Phillips et al., 2003). Clearly, more work needs to be done.

In general, one of the ways by which approaches for the recognition of faces could benefit from the use of video sequences is via the tracking of the face. Tracking for the pose of a face can be used to restrict the search for matches between an image template and a face model across multiple views around expected values, thus reducing the chances of false maxima. In most approaches however, tracking and recognition remain separate processes and the recognition phase usually relies on still images. Relatively few systems have been described that can exploit the temporal continuity and constancy of video sequences. For instance, Li and colleagues (Li, Shaogang Gong, & Heather Liddell, 2001) have described a face recognition system in which the parameters of a 3D point distribution model are estimated using a Kalman filter, effectively tracking parameters and enforcing smoothness over time.

More recently, two approaches have been described that systematically investigate the role of temporal information in video-based face recognition applications. Zhou et al. investigated the recognition of human faces in video sequences using a gallery of both still and video images within a probabilistic framework (Zhou, Krueger, & Chellappa, 2003). In their approach, a time series state-space model is used to extract a spatiotemporal signature and simultaneously characterize the kinematics and identity of a probe video. Recognition then proceeds via marginalization over the motion vector to yield a robust estimate of the posterior distribution of the identity variable using importance-sampling algorithms. Finally, recent work by Zhang and Martinez (2006) convincingly shows that the use of video sequences over still images may help alleviate some of the main problems associated with face recognition (i.e., occlusions, expression and pose changes, as well as errors of localization).

Biological Models for the Perception of Static Faces

Initial biologically inspired computational models for the processing of human faces have focused on the direct implementation of psychological theories. Classically, these theories have assumed abstract cognitive representations such as ''face spaces,'' interpreting faces as points in abstract representation spaces. It has been typically assumed that such points are randomly distributed, for example, with a normal distribution (Valentine, 1991). A central discussion in this context has been whether faces are represented as points in a fixed representation space, which is independent of the class of represented faces (example-based coding), or if faces are encoded in relationship to a norm or average face, which represents the average features of a

large representative set of faces (norm-based or norm-referenced encoding) (Leopold, O'Toole, Vetter, & Blanz, 2001; Rhodes, Brennan, & Carey, 1987; Rhodes & Jeffery, 2006). Several recent experimental studies have tried to differentiate between these two types of encoding (Loffler, Yourganov, Wilkinson, & Wilson, 2005; Rhodes & Jeffery, 2006; Tsao & Freiwald, 2006). Contrary to norm-based representations, which are characterized by a symmetric organization of the face space around a norm face, example-based representations do not assume such a special role for the average face. Refinements of such face space models have been proposed that take into account varying example densities in the underlying pattern spaces. For example, it has been proposed that such density variations could be modeled by assuming a Veronoi tessellation of the high-dimensional space that forms the basis of perceptual judgments (Lewis & Johnston, 1999).

Other models have exploited connectionist architectures in order to account for the recognition and naming of faces (Burton & Bruce, 1993). More recent models work on real pixel images, thus deriving the feature statistics directly from real-world data. A popular approach has been the application of Principal Component Analysis, inspired by the eigenface approach in computer vision (Sirovich & Kirby, 1997; Turk & Pentland, 1991). It has been shown that neural network classifiers based on such eigenfeatures are superior to the direct classification of pixel images (Abdi, Valentin, Edelman, & O'Toole, 1995). At the same time, psychological studies have tried to identify which eigencomponents are relevant for the representation of individual face components (such as gender or race) (e.g., O'Toole, Deffenbacher, Valentin, & Abdi, 1994). More advanced models have applied shape normalization prior to the computation of the eigencomponents (Hancock, Burton, & Bruce, 1996). Such approaches closely resemble methods in computer vision that "vectorize" classes of pictures by establishing correspondences between them automatically and separating shape from texture (e.g., Blanz & Vetter, 1999; Lanitis, Taylor, & Cootes, 1997). Eigenfaces have been combined with neural network architectures, including multilayer perceptrons and radial basis functions (RBF) networks (e.g., Valentin, Abdi, & Edelman, 1997a; Valentin, Abdi, Edelman, & O'Toole, 1997b). Another class of models has been based on the computation of features from Gabor filter responses, including a first filtering stage that is similar to the early processing in the visual cortex (Burton, Bruce, & Hancock, 1999; Dailey & Cottrell, 1999; Dailey, Cottrell, Padgett, & Adolphs, 2002).

Only recently have models been developed that take into account detailed principles derived from the visual cortex. One example is the work by Jiang and colleagues (Jiang et al., 2006) who have applied a physiologically inspired hierarchical model (Riesenhuber & Poggio, 1999) for the position and scale-invariant recognition of shapes to faces, in order to test whether face processing requires the introduction of additional principles compared with the processing of general shapes. This model

also reproduces a variety of electrophysiological results on the tuning of neurons in areas V4 and IT (e.g., Riesenhuber & Poggio, 1999) and results in quantitative predictions that are in good agreement with behavioral and fMRI data (Jiang et al., 2006; Riesenhuber, Jarudi, Gilad, & Sinha, 2004). Our own work discussed in this chapter is based on closely related model architectures.

Biological Models for the Perception of Body Movement

Here we briefly review theoretical models for the recognition of body movements, under the assumption that they might contribute important mechanisms that also apply to the processing of dynamic faces. This idea seems consistent with the fact that face and body-selective regions are often located in close neighborhood in the visual cortex. In monkeys, neurons selective for faces have been found in the superior temporal sulcus and the temporal cortex (e.g., Desimone, Albright, Gross, & Bruce, 1984; Pinsk et al., 2009; Pinsk, DeSimone, Moore, Gross, & Kastner, 2005; Tsao, Freiwald, Knutsen, Mandeville, & Tootell, 2003). (See chapters 8, 9, and 11 for further details.) The same regions contain neurons that are selective for body shapes and movements (Barraclough, Xiao, Oram, & Perrett, 2006; Bruce, Desimone, & Gross, 1986; Oram & Perrett, 1996; Puce & Perrett, 2003; Vangeneugden, Pollick, & Vogels, 2008). Similarly, areas selective for the recognition of faces, bodies, and their movements have been localized in the STS and the temporal cortex of humans, partially in close spatial neighborhood (Grossman & Blake, 2002; Kanwisher, McDermott, & Chun, 1997; Peelen & Downing, 2007; Pinsk et al., 2009, 2005).

To our knowledge, no physiologically plausible models have been developed that account for the properties of neurons that are selective for dynamic face stimuli. In contrast, several exist that try to account for neural mechanisms involved in the processing of dynamic body stimuli (Escobar, Masson, Vieville, & Kornprobst, 2009; Giese & Poggio, 2003; Jhuang, Serre, Wolf, & Poggio, 2007; Lange & Lappe, 2006; Schindler, Van Gool, & de Gelder, 2008).

These models are based on hierarchical neural architectures, including detectors that extract form or motion features from image sequences. Position and scale invariance has been accounted for by pooling neural responses along the hierarchy. It has been shown that such models reproduce several properties of neurons that are selective for body movements and behavioral and brain imaging data (Giese & Poggio, 2003). Recent work proves the high computational performance of biologically inspired architectures for the recognition of body movement, which lies in the range of the best nonbiological algorithms in computer vision (Escobar et al., 2009; Jhuang et al., 2007). Architectures of this type will be proposed in the following discussion as a basic framework for the development of a neural model for the processing of dynamic faces.

A central question in the context of such models has been how form and motion processing contribute to the recognition of body motion. Consistent with experimental

evidence (Casile & Giese, 2005; Thurman & Grossman, 2008; Vangeneugden et al., 2008), some models have proposed an integration of form and motion information, potentially in the STS (Giese & Poggio, 2003; Peuskens, Vanrie, Verfaillie, & Orban, 2005). Conversely, some studies have tried to establish that at least the perception of point-light biological motion is exclusively based on form processing (Lange & Lappe, 2006). Since facial and body motion generate quite different optic flow patterns (e.g., with respect to their smoothness and the occurrence of occlusions), it is not obvious whether the relative influences of form and motion are similar for the processing of dynamic faces and bodies. The study of the relative influences of form and motion in the processing of face stimuli is thus an interesting problem, which relates also to the question of how different aspects of faces, such as static versus changeable aspects (identity versus facial expression), are processed by different cortical subsystems (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000).

Another neural system for the processing of body movement has been found in the parietal and premotor cortex on macaque monkeys (e.g., Fogassi et al., 2005; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fogassi, & Gallese, 2001). A particularity of these areas is that they contain mirror neurons that respond, not only during visual stimulation, but also during the execution of motor actions. An equivalent of the mirror neuron system has also been described in humans (Binkofski & Buccino, 2006; Decety & Grezes, 1999). This observation has stimulated an extensive discussion in cognitive and computational neuroscience as well as robotics and even philosophy. A central question is how the recognition of actions, and especially imitatible actions (Wilson & Knoblich, 2005), might benefit from the use of motor representations. An important hypothesis in this context is that the visual recognition of actions might be accomplished by an internal simulation of the underlying motor behavior (Prinz, 1997; Rizzolatti et al., 2001). A number of computational models in robotics and neuroscience have tried to implement this principle (e.g., Erlhagen, Mukovskiy, & Bicho, 2006; Miall, 2003; Oztop, Kawato, & Arbib, 2006; Wolpert, Doya, & Kawato, 2003).

It has been proposed that a similar process—the internal simulation of somatovisceral states—and potentially even motor commands might also be involved in the recognition of emotional facial expressions (e.g., van der Gaag, Minderaa, & Keysers, 2007). A close interaction between perceptual and motor representations of facial movements is also suggested by the phenomenon of facial mimicry, i.e., the stimulation of electric muscle responses by the observation of emotional pictures of faces (see chapter 11). From a theoretical point of view, these observations raise a question about the exact nature of this internal simulation: Does it, for example, reflect the spatial and temporal structure of facial actions or is it more abstract, e.g., in terms of emotional states?

Basic Neural Architecture

In this section we present a basic neural architecture that is consistent with many experimental results on the recognition of shapes and motion patterns, and even of static pictures of faces. The underlying model formalizes common knowledge about crucial properties of neurons on different levels of the ventral and dorsal visual pathway (Giese & Poggio, 2003; Jiang et al., 2006; Riesenhuber & Poggio, 1999). In addition, architectures of this type have been tested successfully with real-world form and motion stimuli (Jhuang et al., 2007; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). This makes them interesting as a basis for the development of biological models for the processing of dynamic faces.

Feedforward Hierarchical Models of the Ventral Stream of the Visual Cortex

The processing of shape information in the cortex is thought to be mediated by the ventral visual pathway running from V1 (Hubel & Wiesel, 1968) through extrastriate visual areas V2 and V4 to the IT (Perrett & Oram, 1993; Tanaka, 1996), and then to the prefrontal cortex (PFC), which is involved in linking perception to memory and action. Over the past decade, a number of physiological studies in nonhuman primates have established several basic facts about the cortical mechanisms of object and face recognition (see also chapters 7 and 8). The accumulated evidence points to several key features of the ventral pathway. Along the hierarchy from V1 to the IT, there is an increase in invariance with respect to changes in position and scale, and in parallel, an increase in receptive field size and the complexity of the optimal stimuli for the neurons (Logothetis & Sheinberg, 1996; Perrett & Oram, 1993).

One of the first feedforward models for object recognition, Fukushima's Neocognitron (Fukushima, 1980), constructed invariant object representations using a hierarchy of stages by progressively integrating convergent inputs from lower levels. Modern feedforward hierarchical models fall into different categories: neurobiological models (e.g., Mel, 1997; Riesenhuber & Poggio, 1999; Serre, Kreiman et al., 2007; Ullman, Vidal-Naquet, & Sali, 2002; Wallis & Rolls, 1997), conceptual proposals (e.g., Hubel & Wiesel, 1968; Perrett & Oram, 1993), and computer vision systems (e.g., Fukushima, 1980; LeCun, Bottou, Bengio, & Haffner, 1998). These models are simple and direct extensions of the Hubel and Wiesel simple-to-complex cell hierarchy.

One specific implementation of this class of models (Serre et al., 2005; Serre, Kreiman et al., 2007) is the following: The model takes as an input a gray-value image that is first analyzed by a multidimensional array of simple (S1) units which, like cortical simple cells, respond best to oriented bars and edges. The next C1 level corresponds to striate complex cells (Hubel & Wiesel, 1968). Each of the complex C1 units receives the outputs of a group of simple S1 units with the same preferred

orientation (and two opposite phases) but at slightly different positions and sizes (or peak frequencies). The result of the pooling over positions and sizes is that C1 units become insensitive to the location and scale of the stimulus within their receptive fields, which is a hallmark of cortical complex cells. The parameters of the S1 and C1 units were adjusted to match as closely as possible the tuning properties of V1 parafoveal simple and complex cells (receptive field sizes, peak spatial frequency as well as frequency and orientation bandwidth).

Feedforward theories of visual processing, and this model in particular, are based on extending these two classes of simple and complex cells to extrastriate areas. By alternating between S layers of simple units and C layers of complex units, the model achieves a difficult tradeoff between selectivity and invariance. Along the hierarchy, at each S stage, simple units become tuned to features of increasing complexity (e.g., from single oriented bars to combinations of oriented bars to form corners and features of intermediate complexities) by combining afferents of C units with different selectivities (e.g., units tuned to edges at different orientations). For instance, at the S2 level (respectively, S3), units pool the activities of retinotopically organized afferent C1 units (respectively, C2 units) with different orientations (different feature tuning), thus increasing the complexity of the representation from single bars to combinations of oriented bars forming contours or boundary conformations. Conversely, at each C stage, complex units become increasingly tolerant to 2D transformations (position and scale) by combining afferents (S units) with the same selectivity (e.g., a vertical bar) but slightly different positions and scales.

This class of models seems to be qualitatively and quantitatively consistent with (and in some cases actually predicts, several properties of subpopulations of cells in V1, V4, the IT, and the PFC as well as fMRI and psychophysical data. For instance, the described model predicts the maximum computation by a subclass of complex cells in the primary visual cortex (Lampl, Ferster, Poggio, & Riesenhuber, 2004) and area V4 (Gawne & Martin, 2002). It also shows good agreement (Serre et al., 2005) with other data in V4 on the tuning for two-bar stimuli and for boundary conformations (Pasupathy & Connor, 2001; Reynolds, Chelazzi, & Desimone, 1999). The IT-like units of the model exhibit selectivity and invariance that are very similar to those of IT neurons (Hung, Kreiman, Poggio, & DiCarlo, 2005) for the same set of stimuli, and the model helped explain the tradeoff between invariance and selectivity observed in the IT in the presence of clutter (Zoccolan, Kouh, Poggio, & DiCarlo, 2007). Also, the model accurately matches the psychophysical performance of human observers for rapid animal versus nonanimal recognition (Serre, Oliva, & Poggio, 2007), a task that is not likely to be strongly influenced by cortical backprojections. This implies that such models might provide a good approximation of the first few hundred milliseconds of visual shape processing, before eye movements and shifts of attention become activated.

Are Faces a Special Type of Object?

The question of how faces are represented in the cortex has been at the center of an intense debate (see Gauthier & Logothetis, 2000; Tsao & Livingstone, 2008 for recent reviews). Faces are of high ecological significance and it is therefore not surprising that a great deal of neural tissue seems to be selective for faces both in humans (Kanwisher et al., 1997) and in monkeys (Moeller, Freiwald, & Tsao, 2008; Tsao & Livingstone, 2008). On the one hand, electrophysiological studies (Baylis, Rolls, & Leonard, 1985; Perrett, Rolls, & Caan, 1982; Rolls & Tovee, 1995; Young & Yamane, 1992) have suggested that faces, like other objects, are represented by the activity of a sparse population of neurons in the inferotemporal cortex. Conversely, a theme that has pervaded the literature is that faces might be special. For instance, the so-called face inversion effect [i.e., the fact that the inversion of faces affects performance to a much greater extent than inversion of other objects (Carey & Diamond, 1986; Yin, 1969)] suggested that face processing may rely on computational mechanisms such as configurational processing; this would seem incompatible with the shape-based models described earlier, which are based on a loose collection of image features and do not explicitly try to model the geometry of objects.

A model (Riesenhuber & Poggio, 1999) that is closely related to that described earlier was shown to account for both behavioral (Riesenhuber et al., 2004) and imaging data (Jiang et al., 2006) on the processing of still faces in the visual cortex. The model predicts that face discrimination is based on a sparse representation of units selective for face shapes, without the need to postulate additional, "face-specific" mechanisms. In particular, the model was used to derive and test predictions that quantitatively link model FFA face neuron tuning, neural adaptation measured in an fMRI rapid adaptation paradigm, and face discrimination performance. One of the successful predictions of this model is that discrimination performance should become asymptotic as faces become dissimilar enough to activate different neuronal populations. These results are in good agreement with imaging studies that failed to find evidence for configurational mechanisms in the FFA (Yovel & Kanwisher, 2004).

Feedforward Hierarchical Models of the Dorsal Stream

The processing of motion information is typically thought of as being mainly accomplished by the dorsal stream of the visual cortex. Whereas the computational mechanisms of motion integration in lower motion-selective areas (see Born & Bradley, 2005; Smith & Snowden, 1994 for reviews) have been extensively studied, relatively little is known about the processing of information in higher areas of the dorsal stream. It has been proposed that organizational and computational principles may be similar to those observed in the ventral stream (i.e., a gradual increase in the

complexity of the preferred stimulus and invariance properties along the hierarchy) (Essen & Gallant, 1994; Saito, 1993).

Building on these principles, Giese and Poggio have proposed a model for motion recognition that consists of a ventral and a dorsal stream (Giese & Poggio, 2003). Their simulations demonstrated that biological motion and actions can be recognized, in principle, by either stream alone, via the detection of temporal sequences of shapes in the ventral stream of the model, or by recognizing specific complex optic flow patterns that are characteristic for action patterns in the dorsal stream. The architecture of the ventral stream follows closely the architecture of the models described earlier (Riesenhuber & Poggio, 1999), with the addition of a special recurrent neural mechanism on the highest level that makes the neural units' responses selective for sequential temporal order. The dorsal stream applies the same principles to neural detectors for motion patterns with different levels of complexity [such as local and opponent motion in the original model, or complex spatially structured optic flow patterns that are learned from training examples (Jhuang et al., 2007; Sigala, Serre, Poggio, & Giese, 2005]. The model reproduced a variety of experiments (including psychophysical, electrophysiological, and imaging results) on the recognition of biological motion from point-light and full-body stimuli. Subsequent work showed that the dorsal stream is particularly suited for generalization between full-body and point-light stimuli, and produced reasonable recognition results even for stimuli with degraded local motion information, which was previously interpreted as evidence that perception of biological motion is exclusively based on form features (Casile & Giese, 2005).

This line of work has recently been extended by the inclusion of simple learning mechanisms for middle temporal-like units (Jhuang et al., 2007; Sigala et al., 2005), making it possible to adapt the neural detectors in intermediate stages of the model to the statistics of natural video sequences. The validation of this model showed that the resulting architecture was competitive with state-of-the-art computer vision systems for the recognition of human actions. This makes such models interesting for the recognition of other classes of dynamic stimuli, such as dynamic faces.

Extensions of the Basic Architecture for the Processing of Dynamic Faces

The core assumption in this chapter is that the recognition of dynamic facial expressions might exploit computational mechanisms similar to those used to process static objects or body movements. This does not necessarily imply that the underlying neural structures are shared, even though such sharing seems likely with respect to lower and midlevel visual processing. In the following discussion we present a number of extensions of the framework discussed in the preceding section that seem necessary in order to develop models for the processing of dynamic faces. In particular, we speculate that the processing of dynamic face stimuli involves a complex interaction

between motion cues from the dorsal stream and shape cues from the ventral stream (as discussed in the previous sections of this chapter).

Skeleton Model for the Processing of Dynamic Faces

Following the principles that have been successful in explaining the recognition of static objects and faces as well as dynamic body movements, figure 13.1 provides a sketch of how motion and shape cues may be integrated within a model for the processing of dynamic faces. The model extracts characteristic features of dynamic facial expressions through two hierarchical pathways that extract complex form and motion features. The highest levels of these two streams are defined by complex pattern detectors that have been trained with typical examples of dynamic face patterns. Within this framework it is possible to make the form and also the motion pathway selective for temporal sequences by the introduction of asymmetric lateral connections between high-level units tuned to different keyframes of a face sequence (as described in Giese & Poggio, 2003). Whether such sequence selectivity is critical in the recognition of dynamic faces is still an open question. In the proposed model, the information from the form and motion pathways is integrated at the highest hierarchy level within model units that are selective for dynamic facial expressions. Again, it remains an open question for experimental research to demonstrate the existence of face-selective neurons that are selectively tuned for dynamic aspects (see also chapter 8).

We have conducted preliminary experiments with a part of the proposed architecture using videos of facial expressions as stimuli. Testing the model of the dorsal stream described earlier (Jhuang et al., 2007) on a standard computer vision database (Dollar, Rabaud, Cottrell, & Belongie, 2005) that contains six facial expressions (anger, disgust, fear, joy, sadness, and surprise), we found that a small population of about 500 MT/MST-like motion-sensitive model units were sufficient for a reliable classification of these facial expressions (model performance: 93.0% versus 83.5% for the system by Dollar et al., 2005). These MT/MST-like units combine afferent inputs from "V1" model units that are tuned to different directions of motion. After a brief learning stage using dynamic face sequences, these units become selective for space-time facial features such as the motion of a mouth during a smile or the raising of an eyebrow during surprise. It seems likely that shape cues from the ventral stream would also play a key role, if not even a dominant role, in the processing of dynamic faces (see chapter 4). However, the exact integration of motion and form cues can only be determined from future more detailed experimental evidence.

Extension for Norm-Referenced Encoding

As discussed in the second section of this chapter, several experiments on the processing of static pictures of faces have suggested a relevance of norm-referenced encoding for face processing, and potentially even for the representation of other classes of

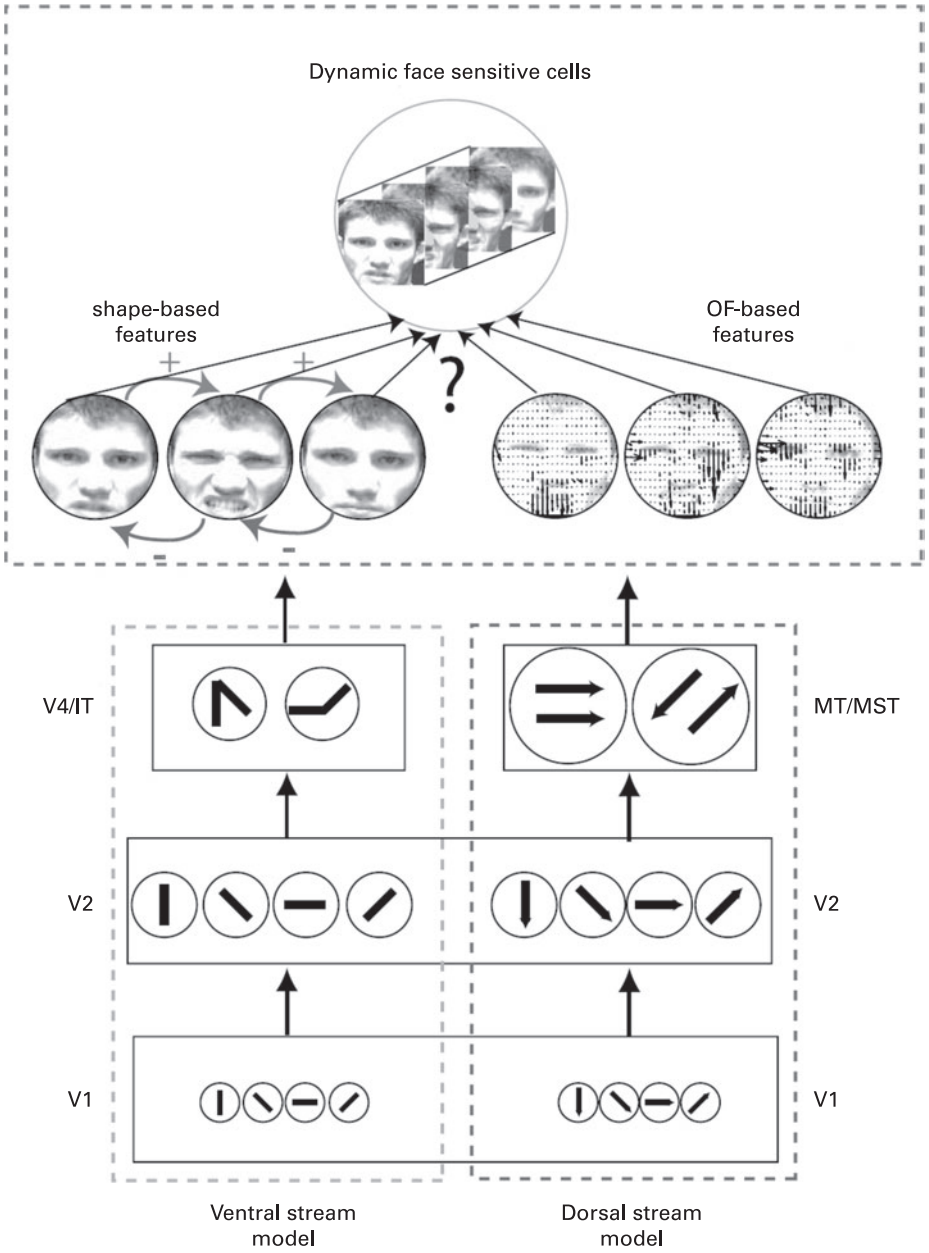


Figure 13.1
Neural model for the processing of dynamic face stimuli. Form and motion features are extracted in two separate pathways. The addition of asymmetric recurrent connections at the top levels makes the units selective for temporal order. The highest level consists of neurons that fuse form and motion information.

objects (Kayaert, Biederman, & Vogels, 2005; Leopold, Bondar, & Giese, 2006; Loffler et al., 2005; Rhodes & Jeffery, 2006; Tsao & Freiwald, 2006). All models discussed so far are example-based. They assume neural units whose tuning depends on the position of a stimulus in feature space, independently of the overall stimulus statistics. An example is units with Gaussian tuning (Riesenhuber & Poggio, 1999; Serre et al., 2005) with centers defined by individual feature vectors that correspond to training patterns.

Conversely, for norm-referenced encoding, the tuning of such units would depend, not only on the actual stimulus, but also on a norm stimulus (average face), resulting in a special symmetry of the tuning about this norm stimulus. The model architecture proposed before can be easily extended to implement ''norm-referenced encoding,'' and it seems that such an extension might be helpful in accounting for the tuning properties of face-selective neurons in the macaque IT area (Giese & Leopold, 2005).

Figure 13.2 shows the results of an electrophysiological experiment (Leopold et al., 2006) that tested the tuning of face-selective neurons in area IT using face stimuli that had been generated by morphing among three-dimensional scans of humans (Blanz & Vetter, 1999). Specifically, morphs between four example faces (F1, F2, F3, and F4) and an average face computed from fifty faces were presented to the animals (panel A). Single-cell and population responses showed a clear tendency of the activity of individual neural units to vary monotonically with the distance of the test stimulus from the average face in the face space (panel B). We then tried to reproduce this result with an example-based model that was basically a simplified version of the model for the ventral stream, as discussed in detail in the previous section. The model consisted of a hierarchy of ''simple'' and ''complex'' units to extract oriented contours and midlevel feature detectors that were optimized by PCA based on the available training patterns. Units at the highest level were Gaussian radial basis functions whose centers were defined by the feature vectors of training faces from the data basis, which is consistent with the example-based models discussed earlier. (See Giese & Leopold, 2005 for further details.)

Although this model achieved robust face recognition with a realistic degree of selectivity, and matching basic statistical parameters of the measured neural responses, it failed to reproduce the monotonic trends of the tuning curves with respect to the distance of the stimuli from the average face that was observed in the experiment (figure 13.2c). This deviation from the data was quite robust against changes in the parameters of the model, or even structural variations like the number of spatial scales. This points toward a fundamental difference with respect to relevant neural encoding principles.

In order to verify this hypothesis, we implemented a second version of the model that included a special neural mechanism that approximates norm-referenced encoding, and which replaced the units with Gaussian tuning in the exemplar-based model.

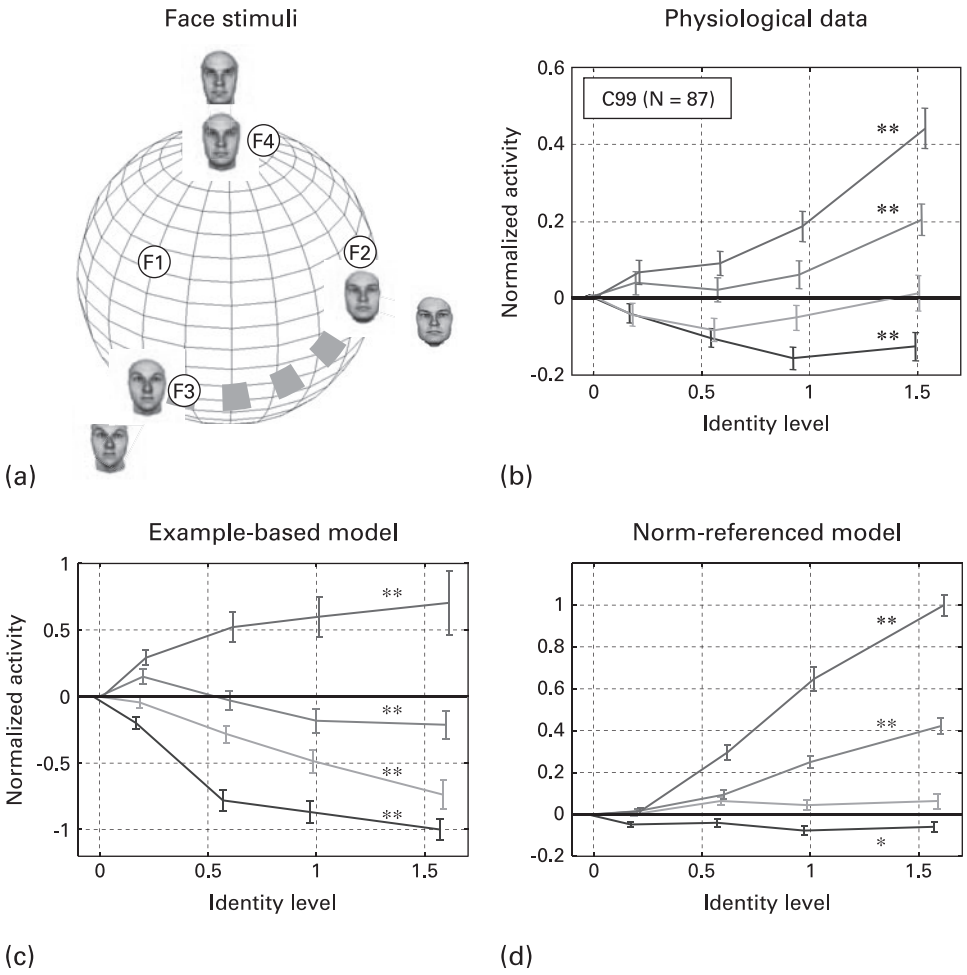


Figure 13.2
Responses of face-selective neurons in area IT and simulation results from two model variants implementing example-based and norm-referenced encoding. (a). Stimuli generated by morphing between the average face and four example faces (F1, F2, F3, and F4). Pictures outside the sphere indicate facial caricatures that exaggerate features of the individual example faces. The identity level specifies the location of face stimuli along the line between the average face and the individual example faces (0 corresponding to the average face and 1 to the original example face). (b). Responses of eighty-seven face-selective neurons (normalized average spike rates within an interval 200–300 ms after stimulus onset) in area IT of a macaque monkey. Different lines indicate the population averages computed separately for different identity levels and for the example face that elicited, respectively, the strongest, second-strongest etc., and the lowest response. Asterisks indicate significant monotonic trends ($p < 0.05$). (c). Normalized responses of face-selective neurons for the model implementing example-based encoding plotted in the same way as the responses of the real neurons in panel b. (d). Normalized responses of the face-selective neurons for the model implementing norm-referenced encoding (conventions as in panel b).

The key idea for the implementation of norm-referenced encoding is to obtain an estimate for the feature vector $\mathbf{u}(t)$ that corresponds to the norm stimulus (average face) by averaging over the stimulus history. Once this estimate has been computed, neural detectors that are selective for the difference between the actual stimulus input from the previous layer $\mathbf{r}(t)$ and this estimate can be easily constructed. The underlying mechanism is schematically illustrated in figure 13.3a.

An estimate of the feature vector $\hat{\mathbf{u}}(t)$ that corresponds to the norm stimulus is computed by "integrator neurons" (light gray in figure 13.3a), which form a (very slow-moving) average of the input signal $\mathbf{r}(t)$ from the previous layer over many stimulus presentations. Simulations showed that for random presentation of stimuli from a fixed set of faces, this temporal average provides a sufficiently accurate estimate of the feature vector that corresponds to the real average face. The (vectorial) difference $\mathbf{z}(t) = \mathbf{r}(t) - \hat{\mathbf{u}}(t)$ between this estimate and the actual stimulus input is computed by a second class of neurons (indicated in white).

The last level of the proposed circuit is given by face-selective neurons (or small networks of neurons) whose input signal is given by the difference vector $\mathbf{z}(t)$ (indicated by dark gray in figure 13.3). The tuning functions of these neurons were given by the function

$$y_k = g_k(\mathbf{z}) \sim |\mathbf{z}| \left(\frac{\mathbf{z}\mathbf{n}_k}{|\mathbf{z}|} + 1 \right)^{\upsilon}, \tag{13.1}$$

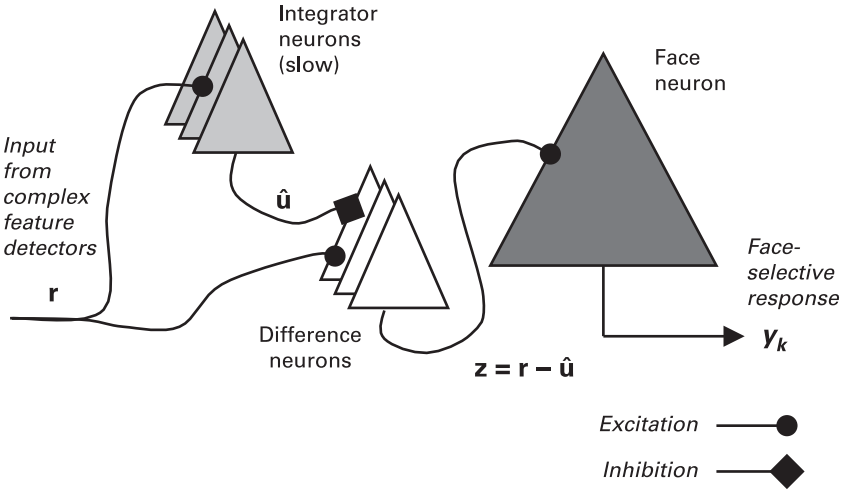
where the first term defines a linear dependence of the output on the length of the distance vector and where the second term can be interpreted as a direction tuning function in the high-dimensional feature space (the unit vector $\mathbf{n}_k$ determining the preferred direction, and the positive parameter υ controlling the width of the direction tuning). While at first glance, this function does not look biologically plausible, it turns out that for $\upsilon = 1$ (a value leading to a good approximation of the physiological data), it can be approximated well by the function

$$y_k = g_k(\mathbf{z}) \sim \mathbf{z}\mathbf{n}_k + |\mathbf{z}| = [\mathbf{z}]_+(\mathbf{n}_k + \mathbf{1}) + [-\mathbf{z}]_+(\mathbf{n}_k - \mathbf{1}). \tag{13.2}$$

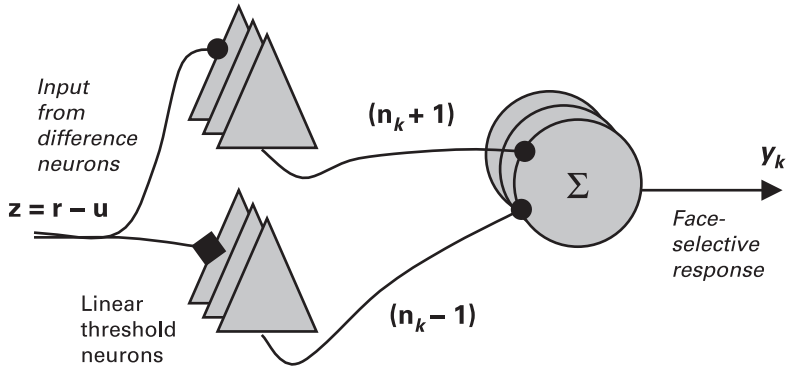
In this formula $[.]_+$ corresponds to a linear threshold function $[x]_+ = \max(x, 0)$, and function (13.2) can be implemented with a simple physiologically plausible two-layer neural network with linear threshold units that is sketched in figure 13.3b.

A quantitative comparison between simulation and real experimental data, using the stimuli and the same type of analysis for the real and the modeled neural data, shows a very good agreement, as shown in figure 13.2d. Specifically, the model reproduces even the number of significant positive and negative trends that were observed in the real data.

This result shows that the proposed architecture can be extended to include norm-referenced encoding without much additional effort and without making biologically



(a)



(b)

Figure 13.3
Neural circuits implementing norm-referenced encoding. (a). Basic circuit deriving an estimate of the feature vector that corresponds to the norm stimulus by averaging and computing the difference between the actual input and this estimate. The difference vector provides the input to the face-selective neurons. (b). Implementation of the tuning function of the face-selective neurons by a two-layer linear threshold network. The unit vectors $\mathbf{n}_k$ define the tuning of the units in the high-dimensional input space (for details, see text).

implausible assumptions. In addition, the proposed circuit can be reinterpreted in a statistical framework as a form of predictive encoding where the face units represent the deviations from the most likely expected stimulus, which is the average face if no further a priori information is given. Predictive coding has been discussed extensively as an important principle for visual processing and especially for object and action recognition (e.g., Friston, 2008; Rao & Ballard, 1997).

Other Missing Components

The model components and principles discussed in this chapter are far from complete, and it seems likely that an architecture that captures all fundamental aspects of the neural processing of dynamic facial expressions will require a variety of additional elements. A few of such principles are listed in the following discussion.

Cortical Feedback The proposed model has primarily a feedforward architecture. It has been shown that in object recognition, such models capture important properties of immediate recognition (in the first 200 ms after stimulus presentation) (Serre, Oliva et al., 2007). For longer stimulus presentations, which is typical for complex dynamic patterns, top-down effects need to be taken into account. This requires the inclusion of top-down connections and attentional effects in the model, which may be particularly important for fine discrimination tasks such as face identification. Overall, the proposed model has been extensively tested for the classification of objects (including faces). Yet vision is much more than categorization because it involves interpreting an image (for faces, this may take the form of inferring the age, gender, and ethnicity of a person, or physical attributes such as attractiveness or social status). It is likely that the feedforward architectures described in this chapter will be insufficient to match the level of performance of human observers on some of these tasks and that cortical feedback and inference mechanisms (Lee & Mumford, 2003) may play a key role.

Attentional Mechanisms Hierarchical architectures that are similar to the proposed model have been extended with circuits for attentional modulation (e.g., Deco & Rolls, 2004; Itti & Koch, 2001). Inclusion of attention in models for dynamic face recognition seems to be crucial since faces have been shown to capture attention (e.g., Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005) and the recognition of facial expressions interacts in a complex way with attention (e.g., Pourtois & Vuilleumier, 2006).

Interaction with Motor Representations and Top-Down Influences of Internal Emotional States These are other potential missing elements. The model described here focuses exclusively on purely visual aspects of facial expressions. The influence of motor

representations could be modeled by a time-synchronized modulation by predictions of sensory states from dynamically evolving predictive internal motor models (e.g., Wolpert et al., 2003). Alternatively, the sensitivity for visual features consistent with specific motor patterns or emotional states might be increased in a less specific manner, similar to attentional modulation without a detailed matching of the temporal structure. Detailed future experiments might help to decide among computational alternatives.

Discussion

In this chapter we have described computational mechanisms that in our view could be important for the processing of dynamic faces in biological systems. Since at present no physiologically plausible model for the processing of such stimuli exists, we have reviewed work from different disciplines: computer vision models for the recognition of dynamic faces (see also chapters 12 and 14), and biologically inspired models for the processing of static faces and full-body movements. In addition, we have presented a physiologically plausible core architecture that has been shown previously to account for many experimental results on the recognition of static objects and faces and dynamic bodies. In addition, our work demonstrates that this architecture reaches performance levels for object and motion recognition that are competitive with state-of-the-art computer vision systems. We suggest that this basic architecture may constitute a starting point for the development of quantitative physiologically inspired models for the recognition of dynamic faces.

As for other work in theoretical neuroscience, the development of successful models for the recognition of dynamic faces will depend critically on the availability of conclusive and constraining experimental data. Although the body of experimental data in this area is continuously growing (as shown by the chapters in the first two parts of this book), the available data are far from sufficient to decide about even the most important computational mechanisms of the processing of dynamic faces. Questions that might be clarified by such experiments include:

- How much overlap is there between cortical areas involved in the processing of static versus dynamic faces?
- How do form and motion cues contribute to the recognition of dynamic faces?
- Is temporal order-selectivity crucial for the processing of facial expressions, and which neural mechanisms implement such sequence selectivity?
- Are neurons tuned to dynamic sequences of face images such as heads rotating in 3D also involved in the problem of (pose) invariant recognition?
- Is there a direct coupling between perceptual and motor representations of facial movements, and what are the neural circuits that implement this coupling?

• How do other modalities, such as auditory or haptic cues, modulate the visual processing of dynamic faces?

The clarification of such questions will likely require the integration of different experimental methods, including psychophysics, functional imaging, lesion studies, and most important, single-cell physiology. An important function for computational models like the ones discussed in this chapter is to quantitatively link the results obtained with various experimental methods and to test the computational feasibility of explanations in the context of real-world stimuli with realistic levels of complexity. Only computational mechanisms that comply with the available data, and which are appropriate for reaching sufficient performance levels with real-world stimuli seem promising as candidates for an explanation of the biological mechanisms that underlie the processing of dynamic faces.

References

Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1995). More about the difference between men and women: Evidence from linear neural networks and the principal-component approach. *Perception, 24*(5), 539–562.

Ahonen, T., Hadid, A., & Pietikaine, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Trans Pattern Anal Machine Intell, 28*(12), 2037–2041.

Barraclough, N. E., Xiao, D., Oram, M. W., & Perrett, D. I. (2006). The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. *Prog Brain Res, 154*, 135–148.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res, 342*(1), 91–102.

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. *Psychonom Bull Rev, 12*, 1048–1053.

Binkofski, F., & Buccino, G. (2006). The role of ventral premotor cortex in action execution and action understanding. *J Physiol Paris, 99*(4–6), 396–405.

Black, M. J., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the fifth international conference on the computer vision (ICCV '95)* (pp. 374–381). Washington DC: IEEE Computer Society.

Blanz, V., & Vetter, T. (1999). A morphable model for synthesis of 3D faces. In *Computer graphics proc. SIGGRAPH* (pp. 187–194). Los Angeles.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annu Rev Neurosci, 28*, 157–189.

Bruce, C. J., Desimone, R., & Gross, C. G. (1986). Both striate cortex and superior colliculus contribute to visual properties of neurons in superior temporal polysensory area of macaque monkey. *J Neurophysiol, 55*, 1057–1075.

Bruce, V., & Young, A. (1986). Understanding face recognition. *Br J Psychol, 77* (Pt. 3), 305–327.

Burton, A. M., & Bruce, V. (1993). Naming faces and naming names: Exploring an interactive activation model of person recognition. *Memory, 1*, 457–480.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognit Sci, 23*(1), 1–31.

Carey, S., & Diamond, R. (1986). Why faces are and are not special: An effect of expertise. *J Exp Psychol Gen, 115*, 107–117.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *J Vis, 5*, 348–360.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. (2003). Facial expression recognition from video sequences: Temporal and static modeling. *Comp Vis Image Understand, 91*(1–2), 160–187.

Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural networks. *Neural Networks, 12*(7–8), 1053–1074.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *J Cognit Neurosci, 14*(8), 1158–1173.

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends Cogn Sci, 3*(5), 172–178.

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res, 44*(6), 621–642.

Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci, 4*(8), 2051–2062.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatiotemporal features. Paper presented at the workshop on visual surveillance and performance evaluation of tracking and surveillance (pp. 65–72). October 16, Beijing China.

Edwards, G., Taylor, C., & Cootes, T. (1998). Interpreting face images using active appearance models. Paper presented at the 3rd IEEE international conference on automatic face and gesture recognition (pp. 300–305). Apr. 14–16 Nara, Japan. Washington DC: IEEE Computer Society.

Erlhagen, W., Mukovskiy, A., & Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Res, 1083*(1), 174–188.

Escobar, M. J., Masson, G. S., Vieville, T., & Kornprobst, P. (2009). Action recognition using a bio-inspired feedforward spiking network. *Int J Comp Vis, 82*(3), 284–301.

Essa, I., Darrell, T., & Pentland, A. (1994). A vision system for observing and extracting facial action parameters. In *Proceedings of the conference on computer vision and pattern recognition (CVPR '94)* (pp. 76–83). 21 Jun–23 Jun 1994, Seattle WA.

Essen, D. C. V., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron, 13*(1), 1–10.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recog, 36*(1), 259–275.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 308*(5722), 662–667.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol, 4*(11), e1000211.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cyb, 36*, 193–202.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119* (Pt 2), 593–609.

Gauthier, I., & Logothetis, N. (2000). Is face recognition not so unique after all? *Cognit Neuropsychol, 17*(1–3), 125–142.

Gawne, T. J., & Martin, J. M. (2002). Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J Neurophysiol, 88*(3), 1128–1135.

Giese, M. A., & Leopold, D. A. (2005). Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing, 65–66*, 93–101.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci, 4*(3), 179–192.

Gong, S., Psarrou, A., Katsoulis, I., & Palavouzis, P. (1994). Tracking and recognition of face sequences. Paper presented at the Proceedings of the European workshop on combined real and synthetic image processing for broadcast and video production. Hamburg, Germany 1994, 23–24. Nov.

Gong, S. M., McKenna, S. J., & Psarrou, A. (2000). *Dynamic vision: From images to face recognition*. London: Imperial College Press.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, 1167–1175.

Hancock, P. J. B., Burton, M. A., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory Cognit, 24*, 26–40.

Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends Cognit Sci, 4*(6), 223–233.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol, 195*(1), 215–243.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast read-out of object identity from macaque inferior temporal cortex. *Science, 310*, 863–866.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nat Rev Neurosci, 2*(3), 194–203.

Jain, A. K., & Li, S. Z. (2005). *Handbook of face recognition*. New York: Springer-Verlag.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *Proceedings of the eleventh IEEE international conference on computer vision (ICCV)* (pp. 1–8). Washington DC: IEEE Computer Society.

Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50*(1), 159–172.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci, 17*(11), 4302–4311.

Kayaert, G., Biederman, I., & Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cereb Cortex, 15*(9), 1308–1321.

Kriegman, D., Yang, M. H., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Trans Pattern Anal Machine Intell, 24*, 34–58.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J Neurophysiol, 92*(5), 2704–2713.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *J Neurosci, 26*(11), 2894–2906.

Lanitis, A., Taylor, C., & Cootes, T. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Trans Pattern Anal Machine Intell, 19*, 743–756.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE, 86*, 2278–2324.

Lee, K. C., Ho, J., Yang, M. H., & Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. Paper presented at the Proceeding International Conference Computer Vision and Pattern Recognition (CVPR '03) (Vol. 1, pp. 313–320). Madison WI, Jun 18–20.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, A 20*(7), 1434–1448.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*(7102), 572–575.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci, 4*(1), 89–94.

Lewis, M. B., & Johnston, R. A. (1999). A unified account of the effects of caricaturing faces. *Vis Cognit, 6*, 1–41.

Li, Y., Shaogang Gong, S., & Heather Liddell, H. (2001). Modelling faces dynamically across views and over time. Paper presented at the 8th IEEE International Conference on Computer Vision (pp. 554–559). Jul 7–14, 2001 Vancouver, BC, Canada.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nat Neurosci, 8*(10), 1386–1391.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annu Rev Neurosci, 19*, 577–621.

Luettin, J., Thacker, N. A., & Beet, S. W. (1996). Visual speech recognition using active shape models and Hidden Markov Models. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 96)* (Vol. 2, pp. 817–820). 07–10 May 1996, Atlanta GA.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Trans, 74*(10), 3474–3483.

Mel, B. W. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp, 9*, 777–804.

Miall, R. C. (2003). Connecting mirror neurons and forward models. *Neuroreport, 14*(17), 2135–2137.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science, 320*(5881), 1355–1359.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Mem Cognit, 22*(2), 208–224.

Oram, M. W., & Perrett, D. I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *J Neurophysiol, 76*(1), 109–129.

Otsuka, T., & Ohya, J. (1996). Recognition of facial expressions using HMM with continuous output probabilities. Paper presented at the 5th IEEE Workshop on Robot and Human Communication (pp. 323–328). Tsukuba, Japan, 11–14 Nov 96.

Oztop, E., Kawato, M., & Arbib, M. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Netw, 19*(3), 254–271.

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans Systems, Man, and Cybernetics, Part B, 36*(2), 433–449.

Pantic, M., & Rothkrant, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Trans Pattern Anal Machine Intell, 22*, 1424–1445.

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *J Neurophysiol, 86*(5), 2505–2519.

Peelen, M. V., & Downing, P. E. (2007). The neural basis of visual body perception. *Nat Rev Neurosci, 8*(8), 636–648.

Perrett, D., & Oram, M. (1993). Neurophysiology of shape processing. *Image Vision Comput, 11*, 317–333.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res, 47*, 329–342.

Peuskens, H., Vanrie, J., Verfaillie, K., & Orban, G. A. (2005). Specificity of regions processing biological motion. *Eur J Neurosci, 21*(10), 2864–2875.

Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., & Bone, M. (2003). Face recognition vendor test 2002. Paper presented at the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (p. 44). 17 Oct 2003 Nice France. Washington DC: IEEE Comp. Society.

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., et al. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative fMRI study. *J Neurophysiol, 101*(5), 2581–2600.

Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proc Natl Acad Sci USA, 102*(19), 6996–7001.

Pourtois, G., & Vuilleumier, P. (2006). Dynamics of emotional effects on spatial attention in the human visual cortex. *Prog Brain Res, 156*, 67–91.

Prinz, W. (1997). Perception and action planning. *Eur J Cogn Psychol, 9*, 129–154.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Phil Trans R Soc Lond B Biol Sci, 358*(1431), 435–445.

Rao, R. P., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex, *Neural Comp, 9*, 721–763.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci, 19*(5), 1736–1753.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cogn Psychol, 19*(4), 473–497.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Res, 46*, 2977–2987.

Riesenhuber, M., Jarudi, I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proc Biol Sci, 271* Suppl *6*, S448–450.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neurosci, 2*(11), 1019–1025.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat Rev Neurosci, 2*(9), 661–670.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol, 73*(2), 713–726.

Rosenblum, M., Yacoob, Y., & Davis, L. (1994). Human emotion recognition from motion using a radial basis function network architecture. In *Proceedings of the 1994 IEEE workshop on motion of non-rigid and articulated objects* (pp. 43–49). Austin TX, Nov 11–12.

Saito, H. (1993). Hierarchical neural analysis of optical flow in the macaque visual pathway. In T. Ono, L. R. Squire, M. E. Raichle, D. I. Perrett, and M. Fukuda (eds.), *Brain mechanisms of perception and memory*. Oxford, UK: Oxford University Press, pp. 121–140.

Schindler, K., Van Gool, L., & de Gelder, B. (2008). Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Netw, 21*(9), 1238–1246.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2005). A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex (AI Memo 2005-036 / CBCL Memo 259). MIT, Cambridge, MA.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Prog Brain Res, 165*, 33–56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA, 104*(15), 6424–6429.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Object recognition with cortex-like mechanisms. *IEEE Trans Pattern Analy Machine Intell, 29*(3), 411–426.

Sigala, R., Serre, T., Poggio, T., & Giese, M. A. (2005). Learning features of intermediate complexity for the recognition of biological motion. *Artificial Neural Networks: Formal Models and Their Applications—ICANN 2005, 15th International Conference, Warsaw, Poland Sep 11–15 2005, Warsaw, Poland* (pp. 241–246).

Sirovich, L., & Kirby, M. (1997). A low-dimensional procedure for identifying human faces, *J Opt Soc Am A, 4*, 519–524.

Smith, A. T., & Snowden, R. J. (1994). *Visual detection of motion*. London: Academic Press.

Suwa, M., Sugie, N., & Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. In *Proceedings of the 4th international joint conference on pattern recognition* (pp. 408–410). Nov 7–10, 1978, New York NY.

Tanaka, K. (1996). Inferotemporal cortex and object vision, *Annu Rev Neurosci, 19*, 109–139.

Thurman, S. M., & Grossman, E. D. (2008). Temporal ''bubbles'' reveal key features for point-light biological motion perception. *J Vis, 8*(3), 1–11.

Tian, Y. L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In S. Z. Li & A. K. Jain (eds.), *Handbook of face recognition.* New York: Springer-Verlag.

Tistarelli, M., Bicego, M., & Grosso, E. (2009). Dynamic face recognition: From human to machine vision. *Image Vis Comp, 27*(3), 222–232.

Tsao, D. Y., & Freiwald, W. A. (2006). What's so special about the average face? *Trends Cognit Sci, 10*, 391–393.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nat Neurosci, 6*(9), 989–995.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annu Rev Neurosci, 31*, 411–437.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition, *J Cognit Neurosci, 3*, 71–86.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat Neurosci, 5*(7), 682–687.

Valentin, D., Abdi, H., & Edelman, B. (1997a). What represents a face? A computational approach for the integration of physiological and psychological data. *Perception, 26*(10), 1271–1288.

Valentin, D., Abdi, H., Edelman, B., & O'Toole, A. J. (1997b). Principal component and neural network analyses of face images: What can be generalized in gender classification? *J Math Psychol, 41*(4), 398–413.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quart J Exp Psychol, 43A*, 161–204.

van der Gaag, C., Minderaa, R. B., & Keysers, C. (2007). Facial expressions: What the mirror neuron system can and cannot tell us. *Soc Neurosci, 2*(3–4), 179–222.

Vangeneugden, J., Pollick, F., & Vogels, R. (2008). Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cereb Cortex, 9*(3), 593–611.

Viola, P., & Jones, M. (2001). Robust real-time face detection. In *Proceedings of the 8th international conference on computer vision* (Vol. 20, No. 11, pp. 1254–1259). Jul 7–14 Vancouver BC, Canada. Washington DC: IEEE Camp. Society.

Wallis, G., & Rolls, E. T. (1997). A model of invariant recognition in the visual system. *Prog Neurobiol, 51*, 167–194.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol Bull, 131*(3), 460–473.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Phil Trans R Soc Lond B Biol Sci, 358*(1431), 593–602.

Yamaguchi, O., Fukui, K., & Maeda, K. (1998). Face recognition using temporal image sequence. In *Proceedings of the 3rd international conference on automatic face and gesture recognition* (pp. 318–323). Nara Japan, Apr 14–16. Washington DC: IEEE Comp. Soc.

Yang, P., Liu, Q., & Metaxas, D. N. (2009). Boosting encoded dynamic features for facial expression recognition. *Patt Recogn Lett, 30*(2), 132–139.

Yin, R. K. (1969). Looking at upside-down faces. *J Exp Psychol, 81*, 141–145.

Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferior temporal cortex. *Science, 256*, 1327–1331.

Yovel, G., & Kanwisher, N. (2004). Face perception: Domain specific, not process specific. *Neuron, 44*(5), 889–898.

Zhang, Y., & Martinez, A. M. (2006). A weighted probabilistic approach to face recognition from multiple images and video sequences. *Image Vis Comput, 24*(6), 626–638.

Zhao, W., Chellappa, R., Rosenfeld, A., & Phillips, P. (2003). Face recognition: A literature survey. *ACM Comp Surveys, 35*(4), 399–458.

Zhou, S., Krueger, V., & Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Comp Vis Image Understand, 91*, 214–245.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J Neurosci, 27*(45), 12292–12307.