Network: Computation in Neural Systems
December 2005; 16(4): 401–418

Taylor & Francis
Taylor & Francis Group

Learning from humans: Computational modeling of face recognition

CHRISTIAN WALLRAVEN, ADRIAN SCHWANINGER, & HEINRICH H. BÜLTHOFF

Max Planck Institute for Biological Cybernetics, Tübingen, Germany

(Received 22 November 2004; revised 25 November 2005; accepted 27 November 2005)

Abstract
In this paper, we propose a computational architecture of face recognition based on evidence from cognitive research. Several recent psychophysical experiments have shown that humans process faces by a combination of configural and component information. Using an appearance-based implementation of this architecture based on low-level features and their spatial relations, we were able to model aspects of human performance found in psychophysical studies. Furthermore, results from additional computational recognition experiments show that our framework is able to achieve excellent recognition performance even under large view rotations. Our interdisciplinary study is an example of how results from cognitive research can be used to construct recognition systems with increased performance. Finally, our modeling results also make new experimental predictions that will be tested in further psychophysical studies, thus effectively closing the loop between psychophysical experimentation and computational modeling.

Keywords: *Face recognition, configural and component information, local features*

Introduction

Faces are one of the most relevant stimulus classes in everyday life. Humans are able to recognize familiar faces with an accuracy of over 90%, even after fifty years (Bahrick et al. 1975). Although faces form a very homogenous visual category in contrast to other object categories, adult observers are able to detect subtle differences between facial parts and their spatial relationship. These evolutionary, very adaptive abilities seem to be severely disrupted if faces are turned upside-down. Consider the classical example shown in Figure 1: it seems that the two faces have a similar facial expression. However, if the two pictures are turned right side up, grotesque differences in the facial expression are revealed (Thompson 1980). In addition, it was shown that—regardless of whether the faces were manipulated or not—inverted faces are much harder to recognize (Yin 1969). As already pointed out by Rock (1973), rotated faces seem to overtax an orientation normalization process making it impossible to succeed in visualizing how all the information contained in a face would look were it to be egocentrically upright. Instead, rotated faces seem to be processed by matching parts, which could be the reason why in Figure 1 the faces look normal when turned upside-down.

Correspondence: Christian Wallraven, Max Planck Institute for Biological Cybernetics, Tübingen, Germany. Tel: +49 7071 601727. Fax: +49 7071 601616. E-mail: christian.wallraven@tuebingen.mpg.de

ISSN: 0954-898X print / ISSN 1361-6536 online
DOI: 10.1080/09548980500508844

Figure 1. Inverted face illusion. When the pictures are viewed right side up (turn page upside-down), the face on the right appears highly grotesque. This strange expression is much less evident when the faces are turned upside-down as depicted here (also known as the "Thatcher illusion" (Thompson 1980)).

In this paper, we will first give a brief overview of several recent psychophysical studies that have been conducted in order to investigate the specific types of information used for face recognition. In accordance with Rock's hypothesis, the results of these studies suggest that faces are processed using two distinct routes: the component route based on detailed, part-based information and the configural route based on geometric information about the layout of these parts. The results of these experiments can be captured in an integrative model that will serve as the basic framework of this paper. In the main part, we will then focus on a specific computational implementation of this framework that will be used to model the behavioral performance in the psychophysical experiments. This implementation uses simple, appearance-based visual features combined with spatial layout information to model the component and configural processing routes. Using the *same* experimental settings as in the psychophysical experiments, we can demonstrate that our straightforward implementation of the two-route processing produces strikingly similar behavior compared to human performance. This shows already that relatively simple image information can support the performance pattern observed in the psychophysical experiments. In a second experiment, we will show that our proposed implementation is not only able to model cognitive results but that it results in increased performance in a difficult recognition task. Specifically, it seems that the configural processing route is able to support recognition across large view angles—a recognition scenario, where artificial recognition systems still fall far behind human performance. Finally, our computational results both in modeling and recognition raise a number of issues that can be verified in further experiments, thus closing the loop between computational and psychophysical research.

Cognitive basis of face recognition

First, we want to briefly discuss what one might call the cognitive basis of face recognition. What information in an image could be used to recognize a given person? In this context, the distinction between parts or component information on the one hand and configural information on the other hand has been used by many studies on human face recognition (for an overview see Schwaninger et al. 2003). In face recognition literature, the term component information (or part-based information) mostly refers to facial elements which are perceived

as distinct parts of the whole such as the eyes, mouth, nose or chin. In contrast, the term configural information refers to the spatial relationship between components and has been used for distances between parts (e.g., inter-eye distance or eye–mouth distance) as well as their relative orientation. There are several lines of evidence in favor of such a *qualitative distinction.* (e.g., Sergent 1984; Tanaka & Farah 1993; Searcy & Bartlett 1996; Schwaninger & Mast 1999; Leder & Bruce 2000; Murray et al. 2000; for a review see Schwaninger et al. 2003). However, one possible caveat of studies that investigated the processing of component and configural information by replacing or altering facial parts is the fact that such manipulations are difficult to separate. Replacing the nose (component change) can alter the distance between the contours of the nose and the mouth, which in turn also changes the configural information. Similarly, moving the eyes apart (configural change) can lead to an increase of the bridge of the nose, which in turn constitutes a component change.

Problems like these were avoided in a recent psychophysical study on face recognition (Schwaninger et al. 2002), which employed a method that did not alter configural or component information, but eliminated either one or the other. The results of two experiments are depicted in Figure 2, where recognition performance is measured in AUC-scores (a psychophysical measure of discriminability measuring the area under the ROC-curve, where $0.5 <= \text{AUC} <= 1.0$, see also Green & Swets 1966).

In the first experiment (black bars in Figure 2), faces were scrambled into their components so that configural information was effectively eliminated. It was found that previously learnt *whole* faces could be recognized by human participants in this scrambled condition (Figure 2, Scr). This result is consistent with the assumption of *explicit representations* of component information in visual memory. In a second condition, a low pass filter that made the scrambled part versions impossible to recognize was determined (Figure 2, ScrBlr). This filter was then applied to *whole* faces to create stimuli in which, by definition, local component-based information would be eliminated. With these stimuli it was then tested

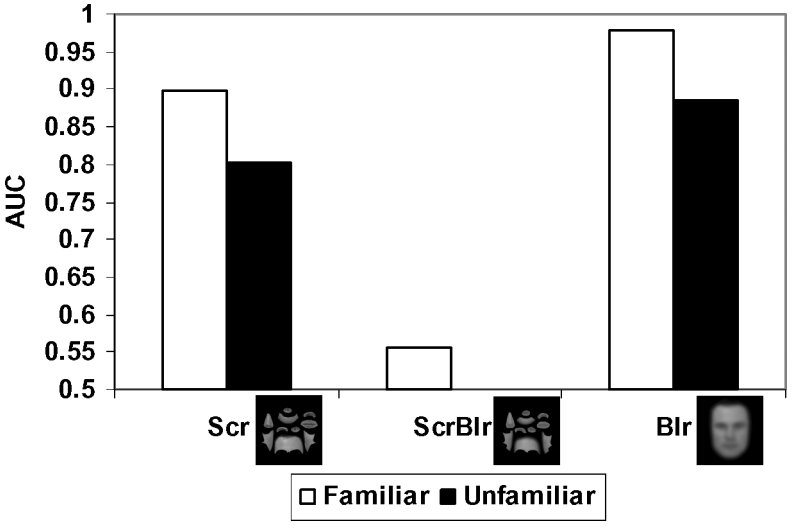

Figure 2. Results from Schwaninger et al. (2002) for familiar and unfamiliar faces, which demonstrate the existence of two separate routes of configural and component processing. The recognition performance of subjects in AUC-values as a function of the three experimental conditions is shown (see text).

whether *configural* information is also explicitly encoded and stored. It was shown that such configural versions of previously learnt faces could be recognized reliably (Figure 2, Blr), suggesting separate explicit representations of both configural and component information. In Experiment 2, these results were replicated for subjects who *knew* the target faces (see white bars in Figure 2). Component and configural recognition results were better when the faces were familiar, but there was *no qualitative* shift in processing strategy since there was no statistical interaction between familiarity and condition. Both experiments provided converging evidence in favor of the view that recognition of familiar and unfamiliar faces relies on component and configural information.

Based on these and other results from psychophysical studies on face processing, Schwaninger et al. (2002; 2003) have proposed an integrative model depicted in Figure 3. In this model, processing of faces entails extracting *local component-based information* and *global configural relations* in order to activate component and configural representations in higher visual areas (so-called face selective areas). The results from the experiments also suggest that the component route relies on detailed image information that is slower to extract than the coarser configural information. Finally, the evidence on familiar face recognition suggests that perceptual expertise simply increases the discriminability for both routes rather than changing the basic structure of the architecture.

The proposed architecture is also compatible with the psychophysical results on the inverted face or Thatcher illusion (Figure 1, Thompson 1980). Inverting the eyes and mouth within an upright face results in a strange activation pattern of component and configural

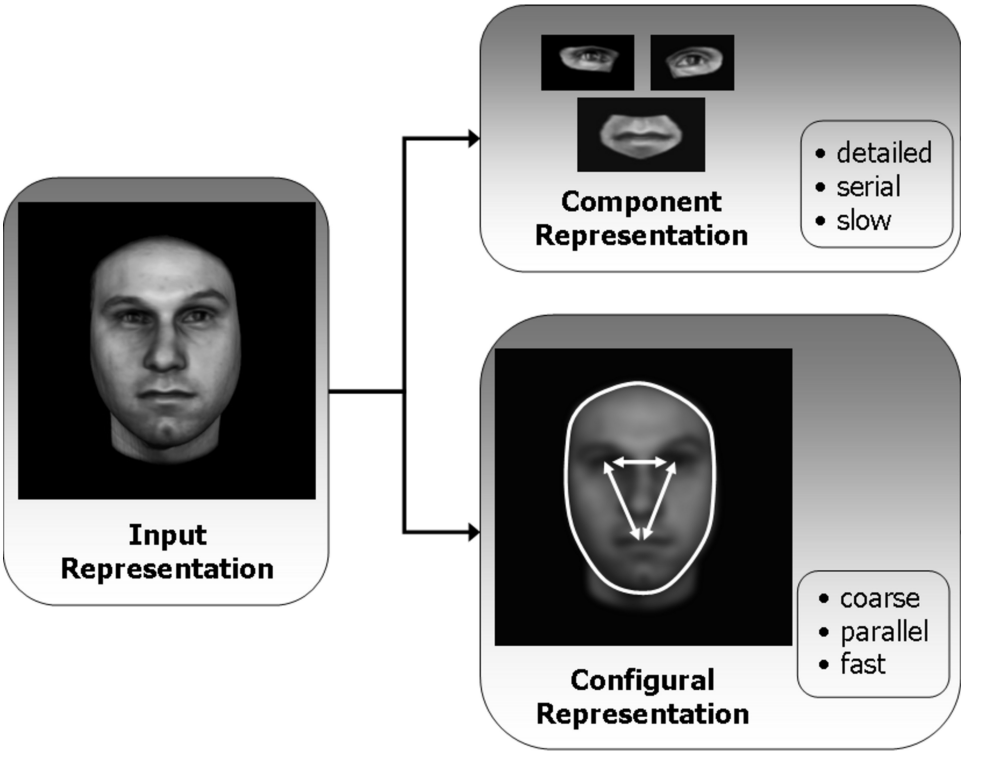

Figure 3. Integrative model for unfamiliar and familiar face recognition showing two distinct processing routes (adapted from Schwaninger et al. 2002).

representations, which consequently results in a bizarre percept. When such a manipulated face is inverted, the activation of configural representations is strongly impaired due to the limited capacity of an orientation normalization mechanism. Consequently, the strange activation pattern of configural representations is reduced and the bizarre percept vanishes. Moreover, in an inverted face the components themselves are in the correct orientation resulting in a relatively normal activation of component representations. Consequently, inverted Thatcher faces appear relatively normal (see also Rock 1988).

Computational approaches to face recognition

Before presenting our implementation of the proposed computational architecture, we want to briefly discuss some of the relevant literature on face recognition in the computational domain. Face recognition is certainly one of the most active topics in computer vision. Within this field, there has been a steady development of algorithms for detection and recognition of faces. Interestingly, computational approaches to face recognition have developed historically from simple, geometric measurements between sparse facial features to appearance-based algorithms working on dense pixel data. Although it was shown that recognition using only geometric information (such as distances between the eyes, the mouth, etc.) was computationally effective and efficient, robust and automatic extraction of facial features has proven to be very difficult under general viewing conditions (see Brunelli & Poggio 1993). In the early 1990s, Turk & Pentland (1991) thus developed a different recognition system called "Eigenfaces", which used the full image information to construct an appearance-based, low-dimensional representation of faces. This approach proved to be very influential for computer vision in general and inspired many subsequent recognition algorithms. Although these algorithms were among the first to work under more natural viewing conditions, they still lacked many important generalization capabilities (including changes in viewpoint, facial expression, illumination and robustness to occlusion). A recent development has therefore been to extract local, appearance-based features, which are more robust to changes in viewing conditions. Some examples of these approaches are: graphs of Gabor Jets (Wiskott et al. 1997), Local Feature Analysis with PCA (Penev & Attik 1996), image fragments (Ullman et al. 2002) and interest-point techniques (Burl et al. 1998; Schiele & Crowley 1999, Wallraven & Bülthoff 2001; Lowe 2004). Going beyond these purely two-dimensional approaches, several approaches have been suggested that use high-level prior knowledge in the form of detailed three-dimensional models in order to provide an extremely well-controlled training set (most notably Blanz et al. 2002; Weyrauch et al. 2004).

Recently, there has been growing interest in testing the biological and behavioral plausibility of some of these approaches (e.g., O'Toole et al. 2000; Furl et al. 2002; Wallraven et al. 2002). However, the work done in this area so far has focused on comparing human performance with a set of "black-box" computational algorithms. Here, we want to go one step further by proposing to look at specific processing strategies employed by humans and trying to model human performance with the help of the integrative model proposed in the previous section. This will not only allow us to better determine and characterize the types of information humans employ for face processing, but also to test the performance of an implementation of the integrative model in other recognition tasks.

Computational implementation

In the following, we describe our implementation of the computational architecture proposed in the previous section. The implementation consists of two main parts: the face

representation which is constructed from an input image, and the matching process which implements the configural and component processing routes.

Face representation

Our computational implementation in this paper is partly inspired by previous studies (Wallraven & Bülthoff 2001; Wallraven et al. 2002) in which an appearance-based computational recognition system based on local features was proposed. This system was shown to provide robust recognition rates in a number of *computational* recognition tasks (Wallraven & Bülthoff 2001) as well as to allow modeling of *psychophysical* results on view-based recognition performance (Wallraven et al. 2002). Here, we develop this system further by including configural and component processing routes as suggested by the psychophysical data discussed earlier.

The algorithm for constructing the face representation proceeds as follows: in a first step, the face image is processed at two scales of a standard Gaussian scale-pyramid to extract localized, visual features. These local features form the basis of our face representation. They are extracted using an interest point detector (in our case a standard "corner detector", Wallraven & Bülthoff 2001), which yields pixel coordinates of salient image regions. Saliency here is defined at the pixel intensity level and can in our case be equated with localized regions in the image that exhibit high curvatures of pixel intensities within their neighborhood. Around each of these located points a small pixel neighborhood is extracted (the size of this neighborhood is 5×5 pixels at each scale) that captures local appearance information. From a computational point of view, this process of feature extraction reduces the amount of storage needed for the face representation significantly—for 50 extracted features from a 256×256 pixel image, the compression rate is 98.1%. In addition, focusing computational resources on salient features also represents an efficient and more robust way of "image" processing. Finally, one can also motivate this choice of feature extraction from both psychological and physiological studies, which support the notion of visual features of intermediate complexity in higher brain areas (Ullman et al. 2002). In addition to these image fragments, for each feature its spatial *embedding* is determined, which consists of a vector containing pixel distances to a number of neighboring features. This vector of distances is used during the matching stage (see next section) to determine either the component or the configural properties of each feature. In order to facilitate later processing, the extracted distance vectors are sorted in increasing order.

Figure 4 shows a reconstruction of a face from such a feature representation, in which features from coarse scales were resized according to the scale difference and then images from the two scales superimposed starting with the coarsest scale. Two aspects are worth noting here: first, even though our representation is sparse, the reconstruction preserves some of the overall visual impression of the original face. Second, and perhaps more importantly, one can see that the extracted features tend to cluster around important *facial* features. Eyes, mouth and nose result in a much higher density of features than, for example, the forehead or the cheeks.

Two additional properties of the visual features chosen for the face representation are worth mentioning. The first property concerns the *scales* at which features are extracted in our implementation. The frequency range of the features used corresponds to around ten cycles per face width for the coarser scale, and 40 cycles per face width for the finer scale, respectively. Interestingly, a recent study by Goffaux et al. (2005) found low-frequency information around eight cycles per face width to be important for configural processing, whereas high-frequency information above 32 cycles per face was important for processing

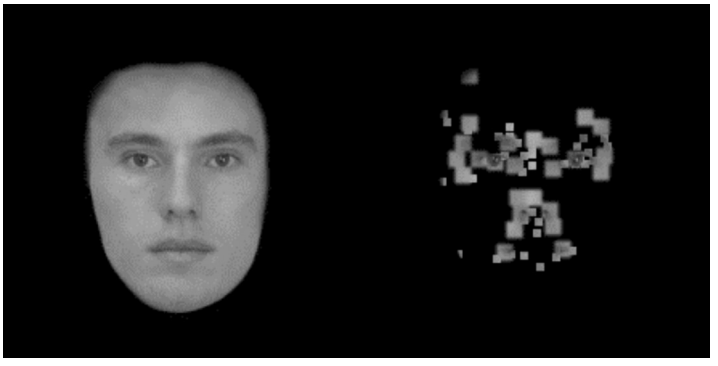

Figure 4. Original face (left) and reconstruction from its feature representation (right). Blurred features originate from the coarse scale, whereas detailed features originate from the fine scale. Note how features tend to cluster around facial landmarks (eyes, nose, mouth).

of facial components. The values of our implementation, which were mainly derived from previous computational experiments (Wallraven et al. 2001; 2002), thus correspond closely to the frequency ranges used by the human visual system to extract the two types of information from a face stimulus. The second property of the extracted visual features concerns their *overlap* across scales. In particular, a study discussed in Maurer et al. (2002) found evidence that information is shared between the configural route and the component route. Although the proposed implementation processes visual information at two different spatial scales, there is some overlap in terms of feature location across scales. This overlap occurs mainly at salient facial features such as the eyes or the corners of the mouth (see Figure 4) and shows that some information is, indeed, shared across scales in our proposed face representation.

It is important to stress that we do not want to claim that a simple scheme like salient feature detection is able to fully explain the complicated processes of feature formation in humans. Indeed, for face recognition there are other types of information available, of which motion—such as resulting from facial expressions or talking—is probably one of the most prominent. It seems, however, that the extracted visual features correspond at least in part to perceptually relevant facial features.

It is this observation which leads to our approach of defining component and configural representations in the context of the two-route architecture proposed in the previous section. First of all, in our implementation we do not use prior knowledge about facial components which would, for example, be available in the form of state-of-the-art facial feature detectors (see Hjelmas & Low 2001 for a recent overview) or through the use of a sophisticated three-dimensional face model (such as the morphable model by Blanz et al. 2002; Weyrauch et al. 2004). Instead the implementation is based on a purely bottom-up, data-driven definition of such "components". In principle, this definition could accommodate a variety of object classes; at the same time, however, it is flexible enough to allow later learning of a more abstract definition of parts.

Components in our framework are defined as *tightly packed conglomerates of visual features at detailed scales* (that is, small clusters of image fragments). This definition captures all the important aspects of the component route in the proposed face recognition architecture without using prior knowledge in the form of pre-learned part models (such as templates for the eyes or spline models for the mouth). It should be noted that components in our framework are defined by a small-scale *configuration* in the feature set. This implies that

processing of components relies on the *relationship between features at detailed scales*. Given the psychophysical evidence and complementing component processing, we can now define configural processing based on the *relationship between features at coarse scales* (that is, large clusters of image fragments). These definitions of the two processing routes will be made explicit in the next section, which deals with *matching* of images.

Component and configural processing

The algorithm for recognition of face images is the second main part of our computational implementation of the proposed architecture. As each image consists of a set of visual features with their embeddings, recognition in our case amounts to finding the *best matching feature set* between a test image and all training images. The two routes for face processing are implemented with two different matching algorithms based on configural and component information, which are derived from a common framework.

Matching of two feature sets is done by an algorithm inspired by Pilu (1997). First, a similarity matrix $\mathbf{A}$ is constructed between the two sets, where each term A_{ij} in the matrix is determined as:

$$A_{ij} = \exp\left(-\frac{1}{\sigma_{app}^2} app^2(i,j)\right) \cdot \exp\left(-\frac{1}{\sigma_{emb}^2} emb^2(i,j)\right). \tag{1}$$

The first term in Equation 1 specifies the appearance similarity (*app*) of two visual features, whereas the second term determines the geometric similarity between the embeddings of the features (*emb*). Appearance similarity is determined by the normalized grey value cross-correlation between the two pixel patches I and $\mathcal{J}$ (in this case both I and $\mathcal{J}$ consist of a 5×5 pixel neighborhood that was extracted around each interest point), which was shown to give good results in previous studies (Wallraven et al. 2001; 2002):

$$app(i,j) = \frac{\sum_{k \in N} (I(k) - \bar{I}) \cdot (\mathcal{J}(k) - \bar{\mathcal{J}})}{\sum_{k \in N} (I(k) - \bar{I})^2 \cdot \sum_{k \in N} (\mathcal{J}(k) - \bar{\mathcal{J}})^2} \tag{2}$$

where k indexes all N pixels in the image fragment I, $\mathcal{J}$, respectively and $\bar{I}$, $\bar{\mathcal{J}}$ determine their mean.

Embedding similarity is defined by the Euclidean distance between the two distance vectors:

$$emb^2(i,j) = \sum_{1 \le k \le M} (d_i(k) - d_j(k))^2 \tag{3}$$

where $d(k)$ is the vector containing distances to all other features sorted in increasing order and M is a parameter which specifies how many dimensions of this vector will be taken into account (see also previous section).

Component matching is done in our framework by restricting M to the first few elements of the distance vector, thus restricting analysis to *close* conglomerates of features—a *local* analysis. *Configural* matching on the other hand relies on *global* relationships, such that M is restricted to the last elements of the sorted distance vector. The size of M should be small for component matching (in our experiments, we used $M = 5$) and larger for the global configural matching (in our experiments, we used $M = |d|/2$), where the latter focuses the matching on similar configurations of the image fragments in the image. In addition, the parameters σ_{app} and σ_{emb} are used to control the relative importance of the two types of information: $\sigma_{app} > \sigma_{emb}$ for the component route, as detailed appearance information is more

important for this route, whereas $\sigma_{app} < \sigma_{emb}$ for the configural route, as more weight should be given to the global geometric similarity between the two feature sets. The matrix **A** thus captures similarity between two feature sets based on a combination of geometric distance information and pixel-based appearance information.

Corresponding features can now be found with a simple, greedy strategy by looking at the largest elements of **A** both in row and column satisfying $A(i,j) > thresh$ ($0 <= thresh <= 1$, see also Pilu 1997; Wallraven et al. 2001; 2002), which yields a one-to-one mapping of one feature set onto the other. The additional threshold *thresh* is used to introduce a global quality metric for the matched features. The percentage of matches between the two feature sets for the component route *and* the configural route then constitute two matching scores, which averaged together yield the final matching score.

Computational modeling and recognition experiments

In this section, we first describe our computational modeling experiments, where our implementation was applied to the psychophysical experiments from Schwaninger et al. (2002) using the exact same stimuli. Thus far, appearance-based computational systems for face recognition have relied largely on either local or global image statistics with little or no configural information. This first set of experiments therefore provides a validation of the proposed computational system—which combines appearance-based, local information with configural, global information—in terms of its psychophysical plausibility. In particular, we were interested to see whether the straightforward implementation of the two-route architecture would be able to capture the performance pattern for scrambled, blurred and scrambled–blurred stimuli that was observed in the psychophysical experiments of Schwaninger et al. (2002).

In a second set of experiments, the degree to which the two separate routes for recognition would be beneficial for other recognition tasks was investigated. For this, we chose a challenging scenario for any computer vision system: recognition across large changes in viewing angle. This set of computational experiments complements the psychophysical modeling in our investigation of face recognition based on configuration and component processing.

Modeling psychophysical results

For the computational modeling experiments, a total of twenty faces from the psychophysical experiment were used. Apart from the original image, each face was available in three versions: scrambled (Scr), blurred (Blr) and scrambled–blurred (ScrBlr). The experimental protocol was as follows: for each run of the experiment, ten faces were selected as target faces and ten faces as distractors. The ten target faces were learned by first extracting visual features and embedding vectors as outlined in the previous section. In the testing phase, a face image was presented to the system, which again extracted the feature representation and then found the highest matching score among the ten learned faces using the two-route matching procedure. The presented face could be either a target or a distractor in one of the three stimulus versions (Scr, ScrBlr, Blr). In order to get a better statistical sampling, this experiment was repeated ten times, each time with a different set of target and distractor faces.

In a next step, the experimental data were converted into a performance measure that can be directly compared with the psychophysical data. For this, the matching scores were converted into an ROC curve by thresholding the matching scores for the target faces (resulting in hit-rates as a function of the threshold) as well as the matching scores for the distractor

faces (resulting in false-alarm-rates as a function of the threshold). Finally, the area under the ROC-curve was measured, which in this case yields a *non-parametric* measure of recognition performance (again, 0.5<=AUC<=1.0) similar to the one used in the psychophysical experiments.

From the description of the implementation in the previous section, one can see that there are a number of internal parameters that will affect the performance of the system in the various experimental conditions. The first parameter is the number of features, which specifies the complexity of the data representation—for this parameter one might expect that more features lead to better overall performance. The second set of system parameters is given by σ_{app} and σ_{emb}, which control the relative importance of appearance and geometric information. The third parameter is the quality threshold *thresh*—for this parameter it can be expected that with increasing threshold the discriminability of the found matches will also increase. These parameters allow us to characterize the parameters of the system with respect to the human performance data obtained in the psychophysical experiments.

Figure 5 shows the experimental results based on a set of parameters selected to fit the psychophysical results. In order to show the contributions of each processing route, the data are presented separately based on the combined matching score. As one can see, scrambled faces (Scr) were recognized only by the component-based route, whereas blurred faces depicting configural information were well recognized by the configural route (Blr). In accordance with the psychophysical data (see Figure 2), both types of computational processing broke down in the scrambled–blurred condition (ScrBlr). Note that although the blurring level was determined using *psychophysical* experiments, image information could also support neither detailed component-based nor global configural analysis for the *computational* system. In addition, a significant advantage of configural over component-based processing was found, which is again consistent with the psychophysical results. These results demonstrate that our framework (with the chosen parameters) seems to be able to capture the characteristics of the two separate routes as found in the psychophysical experiments.

In Figure 6, an example of feature matching in each of the three conditions is given. The component route is active for the scrambled condition, the configural route for the blurred

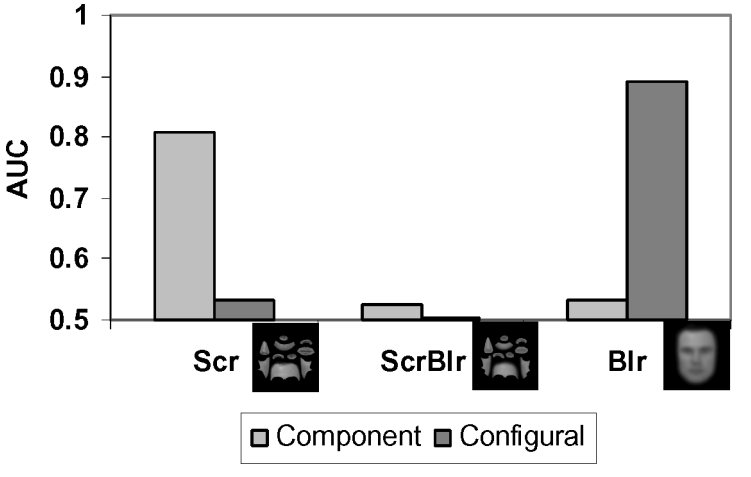

Figure 5. Computational modeling results for unfamiliar face recognition (twenty features in the component route, 25 features in the configural route, *thresh* = 0.925). Shown are the AUC-values for the two processing routes in the three experimental conditions.

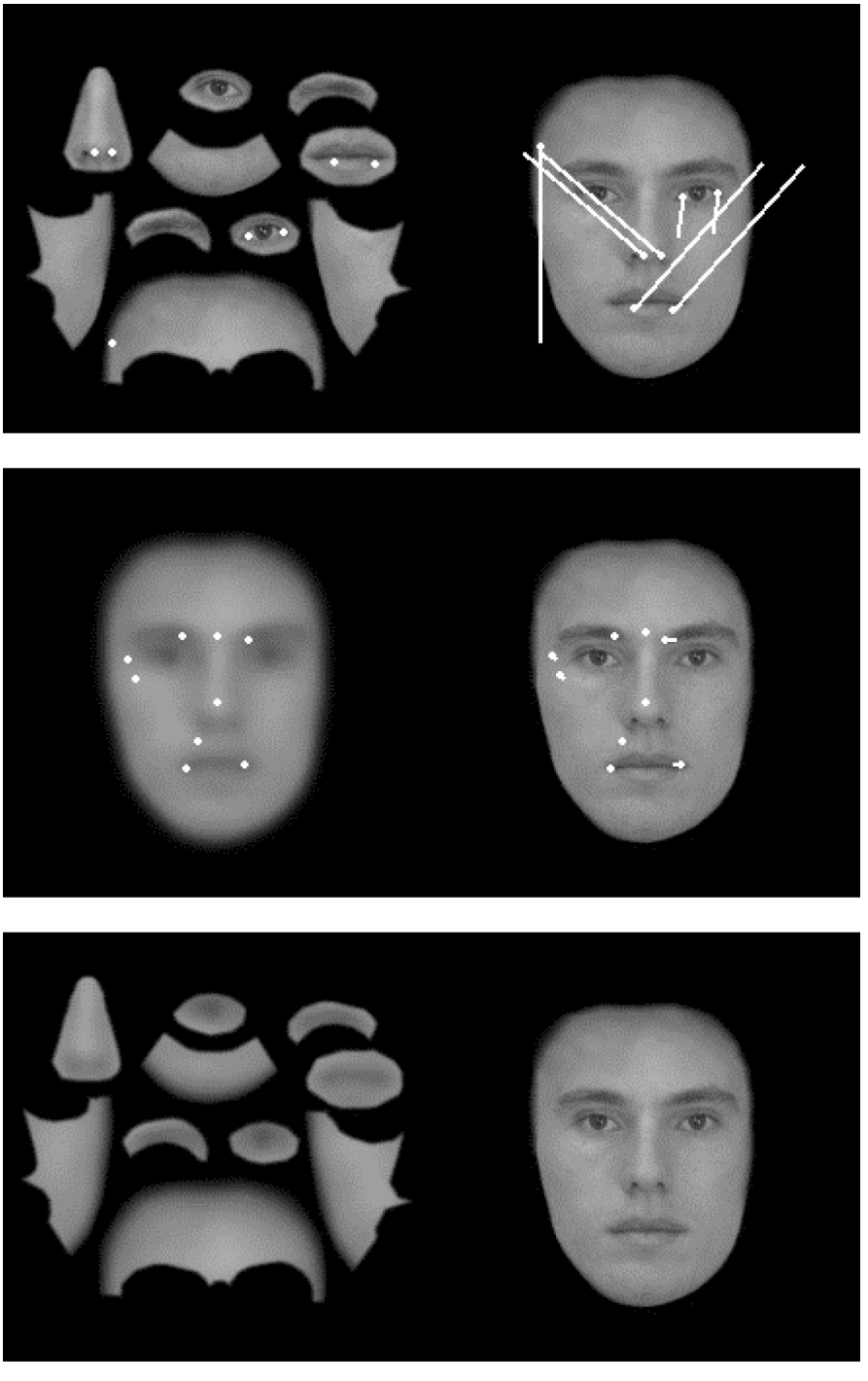

Figure 6. Corresponding features for the three test conditions (upper row: scrambled, middle row: blurred, lower row: scrambled and blurred). Features in the original face (right column) are shown with lines connecting them to the *image location* of their corresponding features in the different conditions (left column). In the scrambled condition, the only active route was the component route, whereas in the blurred condition only configural matches were found. None of the routes were active in the scrambled and blurred condition for this face.

condition and none of the routes for the scrambled and blurred condition. The full experimental results in Figure 5 confirm that both routes process the information *independently* as AUC-values are negligibly small for the conditions in which only one type of information should be present. In addition to the quantitative results and the relative activation of the two routes in the different condition this provides further evidence for the plausibility of the implementation.

Furthermore, Figure 6 shows that component-based matching concentrates on high-level details such as corners of the mouth, points on the nose, some features in the eyes as well as on the eyebrow, etc. Interestingly, this observation already leads to concrete experimental predictions, which can be used to design further psychophysical studies: most of the matching features in the component-based routes focus on high-contrast regions (due to the nature of our visual features). If component-based processing in humans relies on similar low-level information, parts with less high-contrast regions (such as the forehead or the cheeks) should contribute less to the human recognition score. Extending the research of Goffaux et al. (2005), we are currently designing a set of psychophysical experiments which directly address this question of how different parts are weighted in recognition. This represents a good example of how computational modeling feeds back into cognitive research.

Configural matches on the other hand are spread much further apart in the image (tip of the nose, nose bridge, features on the cheek), which carry less appearance information but are globally consistent local features in terms of their spatial layout in the face. The fact that configural processing is largely based on spatial properties also allows for *categorization* of faces, as the configural information captures the global layout of face structure. Figure 7 shows an example of the full matching result between two faces, which demonstrates the generalization capabilities of our system. Again, the only active route in this picture is the configural route—no matches from the highly detailed appearance route were found in this image.

So far, computational modeling has focused on just one example of recognition performance with a selected set of parameters that can reproduce human behavioral results in the unfamiliar condition. In order to strengthen the general assumptions behind our implementation of the processing architecture, it needs to be demonstrated that the same implementation is also able to model the results in the familiar condition *without* changing the relative weights of the processing routes (see discussion in previous section). As mentioned

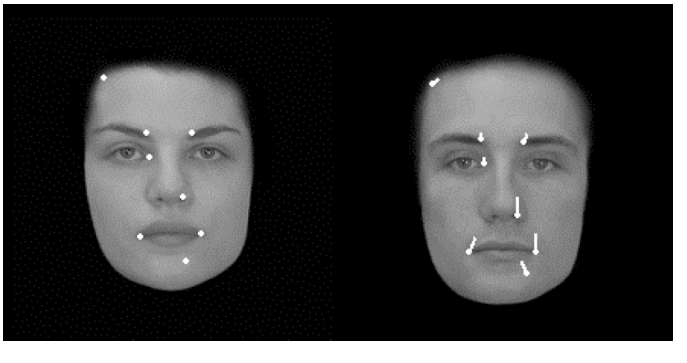

Figure 7. Corresponding features for two faces showing that the general class-based similarity in layout is captured well by the configural route in our implementation. Features in the right face are shown with lines connecting them to the *image location* of their corresponding features in the left face. All corresponding features are found using the configural route only.

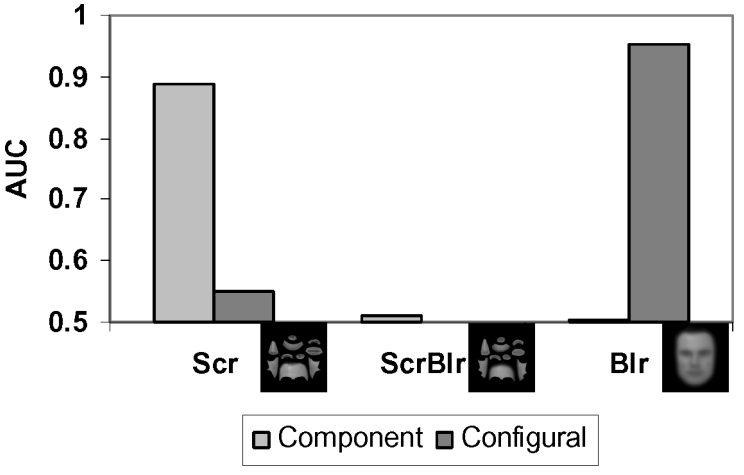

Figure 8. Computational modeling results for familiar face recognition (30 features in the component route, 35 features in the configural route, *thresh* = 0.925).

before, two important parameters in this respect are the number of features and the quality threshold. Both parameters should directly influence the recognition performance when changed—increasing the number of features should allow better generalization, whereas increasing the threshold should result in increased discriminability, and vice versa.

Figure 8 shows the modeling results for a higher number of features (all other experimental conditions were the same as in the previous experiment)—by increasing the number of features while keeping the threshold constant, one can achieve very similar results compared to human performance (cf. Figure 2). This result suggests that one way in which the human visual system might get familiar with an object is simply by enriching its visual representation.

Interestingly, a further increase of the threshold (while keeping the number of features constant) does *not* result in increased performance—performance in the configural route actually decreases slightly (AUC = 0.89 in the blurred condition for *thresh* = 0.95), as well as the performance in the component route (AUC = 0.84 in the scrambled condition for *thresh* = 0.95). A closer investigation shows that for a given visual complexity of the data representation (in this case, 30 features for the component route and 35 features for the configural route) there is an upper limit for the quality threshold. Beyond this limit any increase reduces the number of matches too much, which in turn affects the discriminability of the representation. Better performance using very tight thresholds is thus only possible by also increasing the number of features. Lowering the threshold, on the other hand, has the expected effect of decreasing the discriminability of the matching process thereby resulting in lower recognition performance (AUC = 0.93 for the configural route in the blurred condition, AUC = 0.85 for the component route in the scrambled condition, *thresh* = 0.9).

In summary, our results show that our implementation of the two-route architecture is able to capture the range of human performance observed in the psychophysical experiments. In addition, changes in the internal parameters of the architecture—we have so far investigated visual complexity and discriminability—result in plausible changes in observed performance while retaining the overall qualitative similarity to the human data in terms of the observed weighting of the two routes. Finally, several observations from the computational experiments can be used to plan further behavioral experiments, which will investigate the *features* of face recognition in the context of the two-route architecture.

Recognition across large changes in viewing angle

So far we have demonstrated that our computational implementation of the two-route architecture seems to be suited to model human performance. In a second series of experiments we were interested to see whether there are also benefits of such an architecture in other recognition tasks.

For this, the performance of the configural and component route for face recognition under large view rotations was investigated. This task continues to be a challenge for computational recognition systems. There is good psychophysical evidence that matching of unfamiliar faces for humans is possible even under viewing changes as large as 90° (for a detailed study, see Troje & Bülthoff 1996). In addition, human recognition performance remains highly view-dependent across different viewing angles—a fact that seems to rule out complex, three-dimensional analysis of faces as this would predict a largely view-invariant recognition performance (Biederman & Gerhardstein 1993; but see also Biederman & Kalocsai 1997). Such a three-dimensional strategy in the form of morphable models (Blanz et al. 2002), however, is currently one of the few methods that is able to generalize across larger viewing angles (and also illumination changes) given only *one* image for training. So far, image-based methods—especially based on local features—have met with limited success in this task.

In the following, three different local feature algorithms were benchmarked in a recognition experiment in order to evaluate their performance under large changes in viewing angle. The first algorithm (Std) consists of a simplified version of the one used so far as it uses the same matching framework but is based on local features *without* embedding vectors. This means that the important ingredient for the configural route is not present and that this version of the algorithm relies solely on appearance information. The second algorithm consists of the previously used implementation of the two-route architecture with component and configural processing. The third algorithm is a state-of-the-art local feature framework based on scale-invariant features (SIFT, Lowe 2004), which was shown to perform very well in a number of object recognition tasks. Local features in this framework consist of scale-invariant, high-dimensional (each feature vector has 128 dimensions) histograms of image gradients at local intensity maxima. The SIFT algorithm is available for download at http://cs.spider.uk.ca/~lowe/ and was used without modification in the following experiment. The database used in the following experiments is based on the MPI human face database (see http://faces.kyb.tuebingen.mpg.de, Troje & Bülthoff 1996). This database is composed of three-dimensional high-resolution laser scans of 100 male and 100 female individuals. Each laser scan contains 72000 three-dimensional point coordinates as well as high-resolution RGB texture values. We chose this particular database, as it represents a good compromise between control over viewing conditions on the one hand and visual complexity and realism on the other hand.

From this database, we generated face images in five different poses: $-90°$, $-45°$, $0°$ (frontal view), $45°$ and $90°$. Each grayscale image had a size of 256×256 pixels with the face rendered on a black background. The images were pre-processed such that the faces in the images have the same mean intensity, same number of pixels and same center of mass. The removal of these obvious cues was done to make recognition of the faces a more difficult task. In the following experiments, we report results from a random subset of 100 faces taken from the processed database. This subset was split into a training and test set each containing 50 faces, where the training set consisted of the frontal ($0°$) and profile face views ($\pm 90°$) and the test set of the intermediate ($\pm 45°$) views. Similarly to the previous experiment, an old–new recognition task was chosen to benchmark the algorithms. In order to increase the

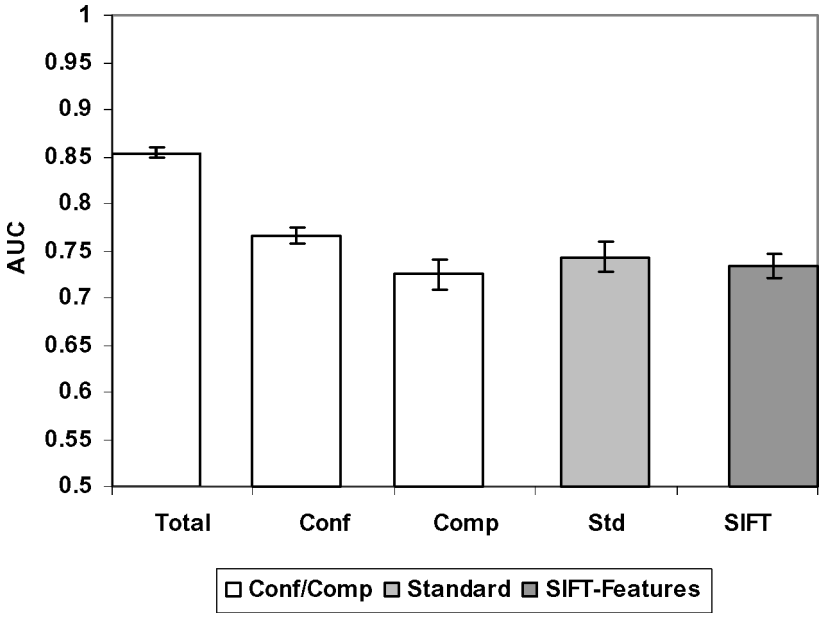

Figure 9. AUC-values for recognition under 45° depth rotation for the proposed two-route architecture (Conf/Comp, showing overall performance of both routes (Total) and the configural (Conf) and component route (Comp)), a simplified version of the algorithm without configural processing (Std) and a state-of-the-art local feature approach (SIFT (Lowe 2004)).

statistical significance, the experiment was repeated ten times with different training and test sets. Figure 9 shows the results as AUC-values (mean and standard deviation) for the three chosen recognition algorithms.

Comparing the overall performance of the two-route algorithm (Figure 9, Total) with the standard (Figure 9, Std) and SIFT framework (Figure 9, SIFT) one can see that the combined processing of configural and component route outperforms both types of algorithms significantly (t-test, $p < 0.001$). Looking closer at the separate results for the configural and component route, the standard algorithm performs only as well as the component route alone ($p = 0.07$, n.s.), even though it operates with the same number of features over the same scales as the combined Total algorithm. Results for the configural route, however, are consistently better than all of the other algorithms and in addition outperform the component route significantly (all $p < 0.001$).

These results show that using rather simple geometric constraints—even in a two-dimensional domain—not only helps to model psychophysical results but also results in increased performance in such a challenging recognition task. Interestingly, the SIFT feature approach does not achieve better results here although it uses much higher-dimensional features than both the two-route processing and the simplified version.[1] The similar performance of both the standard and the SIFT algorithm might show the limitations of a purely pixel-based approach for recognition under such an extreme change in feature appearance. Finally, the results also demonstrate the benefits of *combining* the two different processing routes to form a single recognition output. As this combination yields a much

[1] In principle, it would be possible to add the configural information also to the SIFT approach in order to investigate the performance benefits of spatial constraints in combination with the SIFT features.

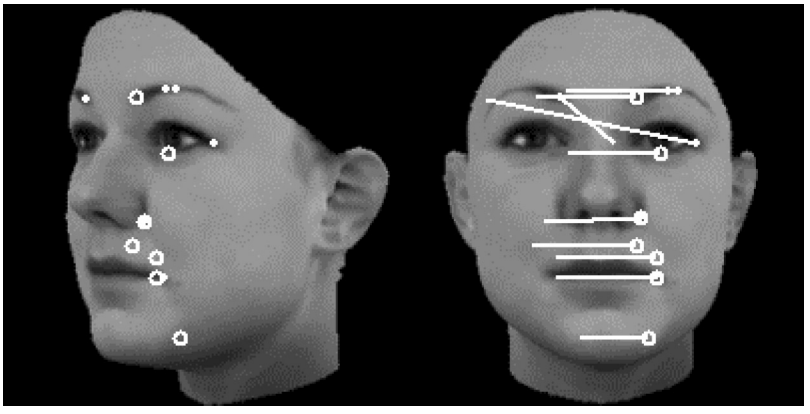

Figure 10. Feature matches under large view rotations—shown is the complete output from both the configural route (shown as circles) and the component route (shown as points). Features in the frontal face (right) are shown with lines connecting them to the *image location* of their corresponding features in the rotated face (left).

higher recognition performance than each of the single routes, this provides further evidence for two distinct and largely independent sources of information which lead to increased discriminability when integrated.

Figure 10 shows an example of matched features—of the eleven matches in total, seven originate from the component route and four from the configural route. The two false matches are component matches, which is not surprising as the probability of false matches increases with increasing viewing angle. Again, component matches focus on smaller details (eyebrows, a part of the eye) whereas configural matches are spread throughout the face.

It has to be mentioned that at the current stage of development it does not seem possible to extend this local feature approach to recognition across larger changes in viewing angle (our experiments show that *all* algorithms break down at 60°, effectively reaching chance level, i.e., AUC = 0.5) as both feature extraction and appearance-based matching component are not robust enough. With regard to further psychophysical experiments, however, our results could represent a first step towards a detailed investigation on what *type* of information might enable humans to generalize across extremely large viewing angles: is configural information more important than component information as our experiments suggest? Are there local features that survive such a large rotation and remain discriminative (blemishes or scars, for example)?

Conclusions

Psychophysical evidence strongly supports the notion that face processing relies on two different routes for configural information and component information. We have implemented a simple computational model of such a processing architecture based on low-level features and their two-dimensional geometric relations, which was able to reproduce the psychophysical results on recognition of scrambled and blurred faces. In particular, we found that in our implementation the configural route was activated in recognizing blurred faces, whereas the component route was activated in recognizing scrambled faces. In addition, the relative recognition performance of these two routes closely matched the human data. In this context it has to be said that an exact *quantitative* modeling—while this might seem a desirable goal—cannot be realistically achieved as there are too many hidden variables in the exact formation of the psychophysical data. Some examples include the different contexts of

human and computational studies: whereas humans have a life-long experience with faces and can use that to encode the faces in both training and test stage, the computational system does not use any prior knowledge about faces. Nevertheless, we were able to reproduce the increase in discriminability seen in the psychophysical data by changing a single parameter in our model—namely the visual complexity of the input representation. Whereas modeling all aspects of familiarity with a stimulus class will certainly entail more than investigating just visual complexity, our results are a good indication that human performance in the psychophysical task and our implementation of the computational architecture might share a similar *functional* (Marr 1982) structure.

Further work in this area will focus on more class-specific processing such as learning of *semantic*, appearance-based parts from local features of faces (see, e.g., Ullman et al. 2002; Weyrauch et al. 2004) together with their spatial layout. In particular, we expect that class-specific information will be important for modeling the upright-advantage in face recognition as well as the Thatcher illusion using our computational framework. The current version of the framework would treat both upright and inverted faces similarly as both feature extraction and configural information are largely invariant to rotation of the face in the image plane. A straightforward extension of our implementation (which we are currently developing) that extracts configural information *only for upright faces*, however, would predict worse performance for inverted faces as the increased discriminability of the configural information is missing. The first step in implementing this extension will consist of the *detection* of upright and inverted faces, which in turn will need to be based on class-specific image information.

With respect to the computational recognition results, one should stress that there are certainly more advanced computational approaches available for both feature representation (Wiskott et al. 1997; Lowe 2004) and matching algorithms (e.g., the SVM framework for local features by Wallraven et al. 2003). Our results using a rather simple set of local features architecture, however, achieved very good recognition performance and should be seen as an attempt to investigate the degree to which the psychophysical data can be explained by *simple, bottom-up strategies*. Taken together, both our modeling and computational results thus demonstrate the advantage of closely coupled computational and psychophysical work in order to investigate cognitive processes.

References

Bahrick HP, Bahrick PO, Wittlinger RP. 1975. Fifty years of memory for names and faces: A cross-sectional approach. J Exp Psychol—General 104:54–75.

Biederman I, Gerhardstein PC. 1993. Recognizing depth-rotated objects—Evidence and conditions for 3-dimensional viewpoint invariance. J Exp Psychol—Hum Perception Perform 19:1162–1182.

Biederman I, Kalocsai P. 1997. Neurocomputational bases of object and face recognition. Phil Trans Royal Soc London Ser B—Biol Sciences 352:1203–1219.

Blanz V, Romdhani S, Vetter T. 2002. Face identification across different poses and illuminations with a 3D morphable model. In Proc. 5th Int Conf on Automatic Face and Gesture Recognition. pp 202–207.

Brunelli R, Poggio T. 1993. Face recognition: Features versus templates. IEEE Trans Pattern Recog Machine Intelligence 15:1042–1062.

Burl MC, Weber M, Perona P. 1998. A probabilistic approach to object recognition using local photometry and global geometry. In Proc ECCV'98. pp 628–641.

Furl N, O'Toole AJ, Phillips PJ. 2002. Face recognition algorithms as models of the other race effect. Cognitive Science 96:1–19.

Goffaux V, Hault B, Michel C, Vuong QC, Rossion B. 2005. The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. Perception 34:77–86.

Green DM, Swets JA. 1966. Signal detection theory and psychophysics. New York: Wiley.

Hjelmas E, Low BK. 2001. Face detection: A survey. Comp Vision Image Understanding 83: 236–274.
Leder H, Bruce V. 2000. When inverted faces are recognized: The role of configural information in face recognition. Quart J Exp Psychol 53A:513–536.
Lowe D. 2004. Distinctive image features from scale-invariant keypoints. Int J Comp Vision 60:91–110.
Marr D. 1982. Vision. San Francisco: Freeman Publishers.
Maurer D, Le Grand R, Mondloch C. 2002. The many faces of configural processing. Trends Cognitive Science 6:255–260.
Murray JE, Yong E, Rhodes G. 2000. Revisiting the perception of upside-down faces. Psychol Science 11:498–502.
O'Toole A, Edelman S, Bülthoff HH. 1998. Stimulus-specific effects in face recognition over changes in viewpoint. Vision Res 38:2351–2363.
Penev PS, Atick JJ. 1996. Local feature analysis: A general statistical theory for object representation. Neural Systems 7:477–500.
Pilu M. 1997. A direct method for stereo correspondence based on singular value decomposition, In Proc Comp Vision Pattern Recognition. pp. 261–266.
Rock I. 1973. Orientation and form. New York: Academic Press.
Rock I. 1988. On Thompson's inverted-face phenomenon (Research Note). Perception 17:815–817.
Schiele B, Crowley JL. 1996. Object recognition using multidimensional receptive field histograms. In Proc ECCV'96. pp. 610–619.
Schwaninger A, Mast F. 1999. Why is face recognition so orientation-sensitive? Psychophysical evidence for an integrative model. Perception (Suppl.) 28:116.
Schwaninger A, Lobmaier J, Collishaw SM. 2002. Role of featural and configural information in familiar and unfamiliar face recognition. Lecture Notes Comp Science 2525:643–650.
Schwaninger A, Carbon CC, Leder H. 2003. Expert face processing: Specialization and constraints. In: Schwarzer G, Leder H, editors. Development of face processing, Göttingen: Hogrefe.
Searcy JH, Bartlett JC. 1996. Inversion and processing of component and spatial-relational information of faces. J Exp Psychol: Human Perception Performance 22:904–915.
Sergent J. 1984. An investigation into component and configurational processes underlying face recognition. Brit J Psychol 75:221–242.
Tanaka JW, Farah M. 1993. Parts and wholes in face recognition. Quart J Exp Psychol 46:225–245.
Thompson P. 1980. Margaret Thatcher—A new illusion. Perception 9:483–484.
Troje NF, Bülthoff HH. 1996. Face recognition under varying poses: The role of texture and shape. Vision Res 36:1761–1771.
Turk M, Pentland A. 1991. Eigenfaces for recognition. J Cognitive Neuroscience 3:71–86.
Ullman S, Vidal-Naquet M, Sali E. 2002 Visual features of intermediate complexity and their use in classification. Nature Neuroscience 5:682–687.
Wallraven C, Bülthoff HH. 2001. Automatic acquisition of exemplar-based representations for recognition from image sequences. In Proc of CVPR 2001—Workshop on Models vs. Exemplars.
Wallraven C, Schwaninger A, Schuhmacher S, Bülthoff HH. 2002. View-based recognition of faces in man and machine: Re-visiting inter-extra-ortho. Lecture Notes Comp Science 2525:651–660.
Wallraven C, Caputo B, Graf A. 2003. Recognition with local features: the Kernel Recipe. In Proc ICCV 2003. pp. 257–264.
Weyrauch B, Huang J, Heisele B, Blanz V. 2004. Component-based face recognition with 3D morphable models. First IEEE Workshop on Face Processing in Video.
Wiskott L, Fellous J, Krüger N, v. d. Malsburg C. 1997. Face recognition by elastic bunch graph matching. IEEE Trans Pattern Machine Intelligence 19:775–779.
Yin RK. 1969. Looking at upside-down faces. J Exp Psychol 81:141–145.