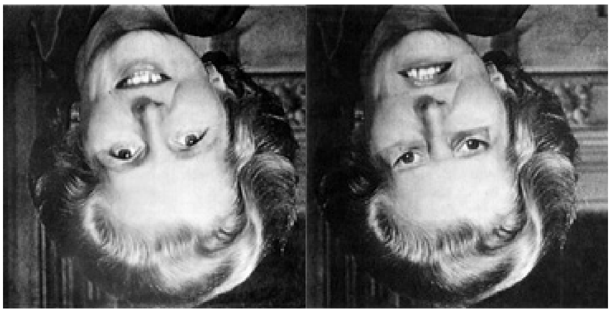
Learning from Humans: Computational Modeling of Face Recognition

Christian Wallraven, Adrian Schwaninger, Heinrich H. Bülthoff
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
christian.wallraven@tuebingen.mpg.de

Abstract: In this paper we propose a computational architecture of face recognition based on evidence from cognitive research. Using an implementation of this architecture based on low-level features and their relations we were able to model aspects of human performance found in psychophysical studies. Furthermore, results from additional recognition experiments show that our framework is able to achieve excellent recognition performance even under large view rotations. Thus, our study is an example of how results from cognitive research can be used to construct recognition systems with better performance. Finally, our results also make new experimental predictions, which can be tested in further psychophysical studies thus closing the loop between experiment and modeling.

1 Cognitive basis of face recognition

Faces are one of the most relevant stimulus classes in everyday life. Although faces in principle are a very homogenous visual category, adult observers are able to detect subtle differences between facial parts and their spatial relationship. Humans are able to recognize familiar faces with an accuracy of over 90%, even after fifty years [1]. These evolutionary very adaptive abilities seem to be remarkably disrupted if faces are turned upside-down. Consider the two pictures in Figure 1: Recognizing the depicted person is more difficult when faces are inverted. Moreover, the two faces seem to have a similar facial expression. Interestingly, if the two pictures are turned right side up, one can easily identify the depicted person and grotesque differences in the facial expression are revealed [9]. As pointed out by Rock [6] 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.



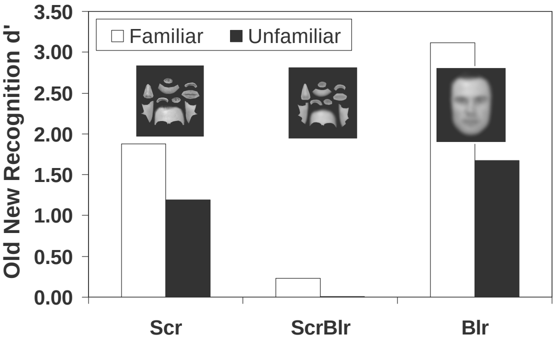
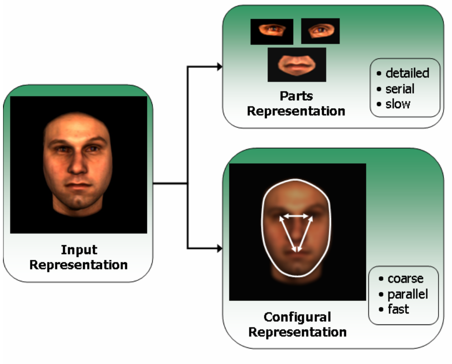


Figure 1 (left). Thatcher illusion. When the pictures are viewed right side up (turn page upside-down), the face on the right appears highly grotesque. This strange expression is much less evident when the faces are turned upside-down as depicted here (from [9]) **Figure 2 (right).** Results from [7] for familiar and unfamiliar faces which demonstrate the existence of two separate routes of configural and component processing (see text).

The distinction between parts or component information on one hand and configural information on the other has been used by many studies on human face recognition (for an overview see [8]). The term component information (or part-based information) has been referred 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 favour such a *qualitative distinction.* 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) sometimes alters the distance between the contours of the nose and the mouth, which in turn also changes the configural information. Similar difficulties apply to configural manipulations (see [8]).

Problems like these were avoided in a recent psychophysical study [7] which employed a method that did not alter configural or component information, but eliminated either the one or the other. The results of two experiments are depicted in Figure 2, where recognition performance is measured in d'-scores. In Experiment 1 it was found that previously learnt faces could be recognized by human participants even when the faces were scrambled into their components so that configural information was effectively eliminated (Figure 2, left). 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, middle). This filter was then applied to *whole* faces to create stimuli in which by definition local part-based information is eliminated in order to test whether configural information is also explicitly encoded and stored. It was shown that configural versions of previously learnt faces could be recognized reliably (Figure 2, right), suggesting separate explicit representations of configural and component information. In Experiment 2 these results were replicated for subjects who *knew* the target faces (white bars in Figure 2). Both experiments provided converging evidence in favour of the view that recognition of familiar and unfamiliar faces relies on component and configural information.



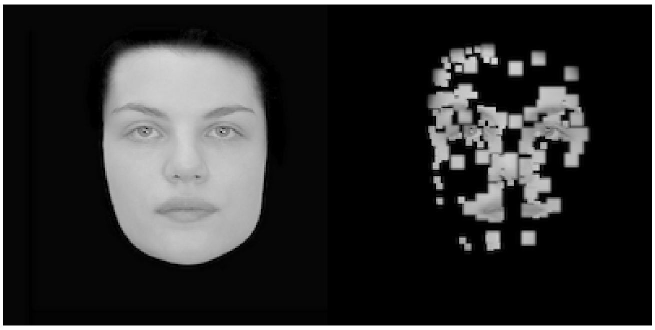


Figure 3 (left). Integrative model for unfamiliar and familiar face recognition. **Figure 4 (right).** Original face (left and reconstruction from its feature representation (right). Blurred features originate from the coarse scale, whereas detailed features originate from the fine scale. Note, how features tend to cluster around facial landmarks (eyes, mouth).

Based on these and other results from psychophysical studies on face processing, Schwaninger et al. [7,8] have proposed an integrative model depicted in Figure 3. In this model, processing of faces entails extracting *local part-based information* and *global configural relations* in order to activate component and configural representations in higher visual areas (so-called face selective areas).

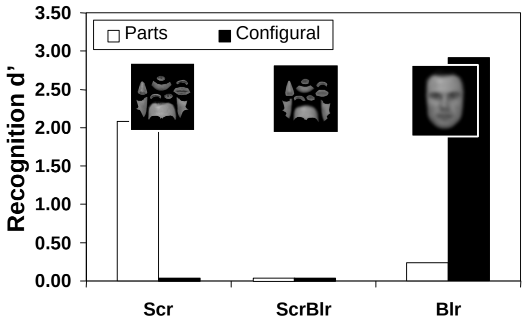
2 Computational modeling

Based on the psychophysical experiments outlined above, we have constructed a computational architecture, which captures the key findings of the model in Figure 3. It is based on previous studies [12,13] in which an earlier version was successfully used to model psychophysical results on view-based object recognition (see also [15] for an approach using similar types of visual information). The algorithm for constructing the face representation (see Fig.4) first extracts low-level visual features at two scales using an interest point detector (such as a standard corner detector [12]), which yields pixel coordinates of salient image regions. Around each point a small neighbourhood of 5x5 pixels is extracted, which captures local appearance information. In addition to these image fragments (see also [11] for another "cognitive" application of such fragments), we also determine for each feature its *embedding*, which consists of a vector containing pixel distances to a number of neighbouring features. In order to facilitate later processing, the extracted distance vectors are sorted in increasing order.

Instead of using prior knowledge about facial parts (see [3] for an overview of feature detectors, or [2] for a 3D morphable face model) we opted for a purely bottom-up definition of such "parts", which can accommodate different object classes but at the same time is flexible enough to allow later learning of a more abstract part definition. Parts in our framework are defined as *tightly packed conglomerates of visual features at detailed scales.* Thus, they are determined by a form of *configuration* in the feature set implying that part-based processing relies on the *relationship between features at detailed scales*. Complementing this, we can now define configural processing (see Figure 3) as the *relationship between features at coarse scales*. An important aspect of our definition of part-based processing is that it does *not* include an explicit clustering of the visual features into semantic parts. Rather, the psychophysically found characteristics of part-based processing will be made explicit during *matching*. As each image consists of a set of visual features, 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 reflected by two types of matching algorithms based on configural and part information. Matching of two feature sets is done by first constructing a similarity matrix **A** [5] between the two sets:

$$A_{ij} = \exp(-\frac{1}{s_{app}^{2}} NCC^{2}(i,j)) \cdot \exp(-\frac{1}{s_{dist}^{2}} c^{2}(i,j)) \qquad (1)$$

The first term specifies an *appearance similarity*, which is determined by the normalized grey value cross-correlation (*NCC*) between the two pixel patches *i* and *j* [12,13]. The second term specifies a two-dimensional *geometric similarity* given by the c^{2} distance between the two distance vectors treated as histograms. *Part-based* matching is now defined by evaluating the c^{2} distance only for the first few elements of the histogram, resulting in a *local* analysis of close conglomerates of features. *Configural* matching on the other hand relies on *global* relationships with the second term in (1) evaluated for the last elements of the sorted distance vectors. **A** thus captures similarity between two feature sets based on a combination of distance and appearance information (s_{app} and s_{dist} can be used to control their relative importance). Corresponding features finally are determined by finding the largest elements of **A** both in row and column with *A(i,j)>thresh* [5,12] yielding a one-to-one mapping between the two feature sets. The percentage of matches for the part-based route *and* the configural route then constitute the two final matching scores for recognition.



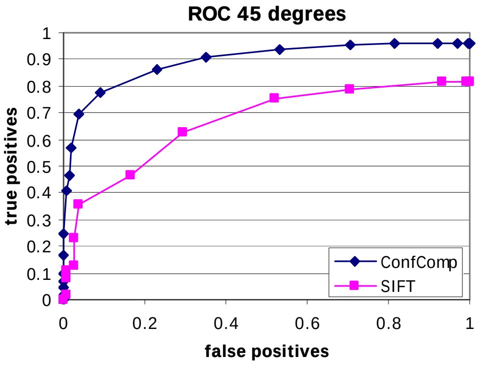


Figure 5 (left). Computational modeling results showing the output of both part and configural route as d'-scores. Note, how the different routes are active in the different conditions. **Figure 6 (right).** ROCs for recognition under 45° depth rotation for the local feature approach (SIFT[4]) versus our approach (ConfComp).

3 Modeling and computational experiments

We applied our implementation outlined above to the psychophysical experiments from section 1 using the exact same stimuli. Figure 5 shows how scrambled faces (Scr) were well recognized by the part-based route whereas blurred faces depicting configural information were well recognized by the configural route (Blr). Perfectly consistent with the psychophysical data (see Figure 2), both types of computational processing break down in the scrambled-blurred condition (ScrBlr), where the information could support neither detailed part-based *nor* global configural analysis. In accordance with the psychophysical study, we also found a significant advantage of configural over part-based processing. These results demonstrate that our framework is able to capture the characteristics of the two separate routes as found in the psychophysical experiments. In addition, we found that part-based recognition concentrates on high-contrast details such as corners of the mouth and eyes, features on the eyebrow, etc. If part-based processing in humans relies on similar low-level information one should find in a psychophysical experiment that parts with less high-contrast regions (such as the forehead or the cheeks) contribute *less* to the human recognition score. We are currently designing a set of experiments, which directly address the question of how different parts are weighted.

In a second series of experiments we investigated recognition of faces under large view rotations. Figure 6 shows ROCs for a recognition experiment with 100 faces, where the training set consisted of frontal face views and the test set of ±45° views. In this fairly challenging recognition experiment, we benchmarked our proposed framework against a recently developed local feature algorithm based on scale invariant features (SIFT), which was shown to provide excellent recognition results in a number of tasks [4]. Even though both our image representation and matching schemes are less sophisticated, one can see a large performance difference between the two approaches (96% versus 81% peak performance). This effect is an extreme example of the benefits of using configural information, which in this case was the *only* active route. Regardless of the computational advantage presented here, our results represent a first step towards a detailed investigation on what *type* of information might enable humans to generalize across extremely large viewing angles for face recognition using an appearance-based approach (see [10], where it was shown that generalization from front to profile view (90 degrees) is possible, but see also [2] which used prior knowledge of a 3D morphable model).

5 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 model the psychophysical results in a qualitative manner. 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. A qualitative similarity is, however, a good indication that the basic assumptions of the architecture and its implementation share a similar structure. One should also add that there are certainly more advanced computational approaches available for both feature representation ([4,15], one could also add configural information to the SIFT features) and matching algorithms (e.g., the SVM framework for local features by [14]). On the other hand, our recognition results using our simple architecture already achieved very good performance levels. 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 (of vision).

References

[1] H.P. Bahrick, P.O. Bahrick and R.P. Wittlinger. Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General,* 104, 1975.

[2] V. Blanz, S. Romdhani and T. Vetter. Face Identification across Different Poses and Illuminations with a 3D Morphable Model. In *Proc. 5th Int. Conference on Automatic Face and Gesture Recognition*, 2002.

[3] E. Hjelmas and B.K. Low. Face Detection: A Survey. *Computer Vision and Image Understanding,* 83. 2001.

[4] D. Lowe. Object recognition from local scale invariant features. *Proc. ICCV'99*. 1999.

[5] M. Pilu. A direct method for stereo correspondence based on SVD. *Proc. CVPR'97*, 1997.

[6] I. Rock. *Orientation and form*. New York: Academic Press. 1973.

[7] A. Schwaninger, J. Lobmaier, and S.M. Collishaw. Role of featural and configural information in familiar and unfamiliar face recognition. In *Proceedings Biologically Motivated Computer Vision.* 2002.

[8] A. Schwaninger, C.C. Carbon, and H. Leder. Expert face processing: Specialization and constraints. In G. Schwarzer and H. Leder, *Development of face processing*, Göttingen: Hogrefe. 2003.

[9] P. Thompson. Margaret Thatcher -- A new illusion. *Perception*, 9(4), 1980.

[10] N.F. Troje and H.H. Bülthoff: Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36, 1996.

[11] S. Ullman, M. Vidal-Naquet and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7), 2001.

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

[13] C. Wallraven, A. Schwaninger, S. Schuhmacher and H.H. Bülthoff. View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho. In *Proc. BMCV'02,* 2002.

[14] C. Wallraven, B. Caputo and A. Graf. Recognition with Local Features: the Kernel Recipe. *Proc. ICCV'03*. 2003.

[15] L. Wiskott, J. Fellous, N. Krüger and C. v. d. Malsburg. Face Recognition by Elastic Bunch Graph Matching. IEEE PAMI 19(7). 1997.