From: Michael A. Arbib (ed.), The Handbook of Brain Theory and Neural Networks, 2nd edition, pp. 434–437. MIT press, 2002.

Face Recognition, Neurophysiology, and Neural Technology

Rolf P. Würtz

Institut für Neuroinformatik
Ruhr-Universität Bochum
D–44780 Bochum, Germany.
http://www.neuroinformatik.ruhr-uni-bochum.de/PEOPLE/rolf/
Rolf.Wuertz@neuroinformatik.ruhr-uni-bochum.de

The recognition of other individuals is a major prerequisite for human social interaction and hence a rather important brain function. The most prominent cue for that recognition is the face. The capability to recognize persons from their faces is part of a spectrum of related skills, which include face segmentation, i.e., finding faces in a scene or image, the estimation of the pose, the direction of gaze, and the person's emotional state. This article focuses on recognition of identity, a more detailed treatment of the other aspects can be found in [Face Recognition, Psychology, and Connectionism].

Neurophysiology

From neuropsychological studies of patients with brain injuries it is known that there are subsystems in the brain that are specialized in face processing. Brain injury can lead to the loss of the capability of face recognition, a deficit called *prosopagnosia*, while leaving the recognition of general objects intact. The opposite dissociation is reported in (Moscovitch et al., 1997). about a patient with intact face recognition together with highly impaired general object recognition. Various stunning perceptual demonstrations show that faces

are perceived differently when viewed upside down or as photographic negatives. Those image manipulations make little difference for the perception of general objects but can modify the perception of identity and expression considerably. These findings lead to the assumption that different brain circuits are used for the processing of general objects and faces, respectively, but there is also considerable evidence that not only faces receive special treatment but all object classes for which there is high expertise (Gauthier et al., 1999).

Other studies show that prosopagnosia patients without any conscious recognition of facial identity still show an unconscious reaction to familiar faces, which is revealed by changes in skin conductance. This mechanism seems to play a major role in the emotional reaction to facial stimuli.

Single unit recordings in the *inferotemporal cortex* of macaque monkeys have revealed neurons with a high responsiveness to the presence of a face, an individual, or the expression on the face (see (Desimone, 1991) for a review). Although the notion of the optimal stimulus for a cell is very hard to probe experimentally some of these cells are as close to grandmother cells [Associative Networks] as the experimental evidence gets.

In humans, cells that become active when a familiar face is seen have been identified in the *inferotemporal* and the *fusiform gyrus* in both hemispheres. Their clusters do not form anatomically well-defined subregions but are neighbored by modules of different specificity, and their location and extent varies considerably among individuals.

A good account of the current knowledge about face recognition in the human brain is given by a model of (Haxby et al., 2000), which refines a cognitive model by Bruce and Young (see chapter 3 of (Young, 1998)) and attaches anatomical locations to its modules. They propose a *core system* for face processing, which consists of three interconnected modules. The first, located in the *inferotemporal occipital gyrus* is responsible for the early extraction of features relevant for faces. The second, in the *superior temporal sulcus* codes for the changeable properties of faces like direction of gaze, lip movement, expression, etc. Identity as an *invariant* face property is processed in the *lateral fusiform gyrus*. This core system communicates with other parts with a need for facial information like attention modules, auditory cortex, and emotional centers. The essence of face recognition, linking the visual information to a name and biographical knowledge about particular persons is carried out in the *anterior temporal lobe*. These other parts make up the *extended system*.

Computational theory

As for all object recognition, the main problem to be solved by a face recognition procedure is *invariance*. The same face can produce very different images under varying position, pose, illumination, expression, partial occlusion, background, etc. The task of the recognition system is to generalize over all these variations and capture only the identity.

It should be noted that this sort of invariant recognition is a ubiquitous property of natural brains but does not come very naturally in current artificial neural network models. Even the simplest case, invariance under translations in the input plane, is difficult to obtain. One major approach starts at the observation that complex cells generalize about small translations of the signal. This can be iterated and leads to hierarchical networks like the Neocognitron [Neocognitron]. A huge advantage of such purely feedforward networks is their speed of processing.

Very little is known about how invariant recognition can be learned from examples and generalized to other instances. In an abstract sense, the important long-term goal is to teach a network precisely the invariances required for a given problem domain. This is directly relevant for face recognition, because invariance under expression and slight deformations are very difficult to capture analytically.

If the only invariance required is translation, then template matching [Object recognition] can solve the problem rather efficiently. A stored pattern (which we will call "model") is compared to an image by shifting the model across the image and taking the scalar product with appropriate normalization at all possible image locations. The maximum of the resulting matrix can serve as as similarity measure between both images.

In order to extend this method to the more complicated invariances involved in face recognition the notion of a *correspondence map* is helpful. Correspondence, central to many problems in computer vision, can be defined as follows.

Point pairs from two given images of the same face correspond if they originate from the same point on the physical face.

Once these correspondences have been established for sufficiently many points an invariant similarity measure between model and object can be defined as the sum or average over the similarities of local features of all corresponding point pairs. As the points on the real face are not accessible to either the brain or a computer, these correspondences can only be estimated

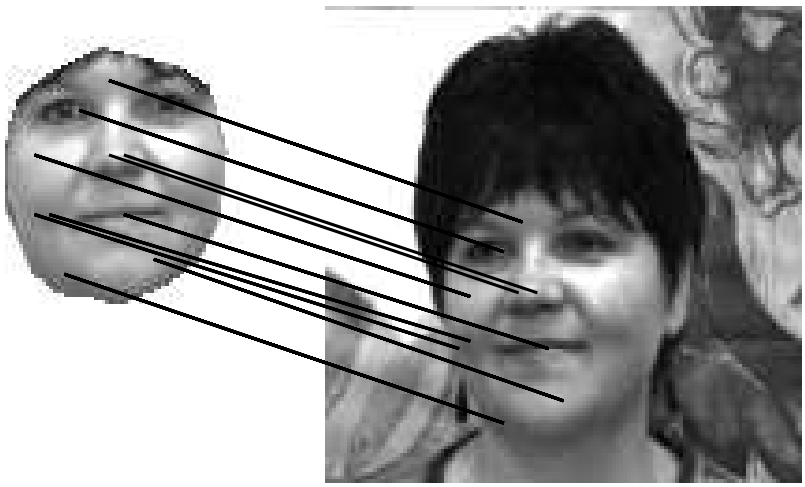

Figure 1: Correspondence maps provide a basis for an invariant similarity measure between two facial images, which can be used for person identification. They also deliver information about pose, size, expression, and are crucial for animation. Their computation is difficult and rarely perfect. The figure shows selected correspondences obtained with the algorithm from (Würtz, 1997)

on the basis of image information. Strictly speaking, correspondences are only defined between images of the same person, but all faces are sufficiently similar in structure to extend the notion to correspondence maps between different faces. These maps have many applications beside recognition, see [Face Recognition, Psychology, and Connectionism].

A system to recognize a person out of a collection of known ones can proceed as follows. Correspondence maps are estimated between the given image and all stored models, similarities are calculated on the basis of the correspondence maps, and the model with the highest similarity is picked as the recognized person. A measure for the *reliability* of the recognition can be derived by a simple statistical analysis of the series of all similarity values.

As correspondence finding is a slow process, the database of known individuals must be organized such that the need for it is minimized. Furthermore, it should not be applied to arbitrary images, but some filtering must select image portions that are likely to contain a face for processing and recognition.

Summarizing the computational theory reveals the following building blocks for a successful face recognition system.

1. A representation of the facial images
2. A method of solving the correspondence problem
3. A similarity measure derived from a pair of images and a correspondence map
4. Organization of the database of known individuals
5. Filtering of the visual data (face finding)

For general reviews of face recognition systems see (Grudin, 2000) and (Chellappa et al., 1995).

Image representation

Many models for face recognition work directly on image grey values or retinal images. In this case, the correspondence problem becomes particularly difficult, as many points from very different locations share the same pixel

value without actually corresponding to each other. A possible remedy consists in combining local patches of pixels. The larger the patch, the more this ambiguity is reduced. On the other hand, the features become more sensitive to distortions and changes in background and are thus of less value for the other required invariances. Patch building may also include linear combinations of pixel values. In this context, *Gabor functions* [Gabor Wavelets for Statistical Pattern Recognition] as a model of simple and complex cells in V1 have turned out to be a good compromise between locality and robustness and are well suited for correspondence finding.

The possibility to process the amplitudes and phases of the Gabor wavelet responses separately is very useful for face processing. Amplitudes (which model the activity of complex cells) vary rather smoothly across the image and so do the similarities of all image features to a single one. Consequently, they provide smooth similarity landscapes well suited for matching templates or single feature vectors. The phases, on the other hand, vary as rapidly as dictated by their center frequency and proceed roughly linearly on image paths in the respective direction. Therefore, they can be used to estimate correspondences with subgrid accuracy (Würtz, 1997; Wiskott et al., 1997).

An important alternative for image representation is to use local features which are derived directly from the statistics of facial images. A prominent example is the neural network based *local feature analysis* (Penev and Atick, 1996), which allows to learn local descriptors by minimizing their correlation. This results in a sparse code adapted for the class represented by the training examples.

Correspondence finding

The representation of a face in terms of local features serves two purposes. First, correspondences must be estimated on the basis of feature similarity, and second, the feature similarities constitute the image similarity. In principle, different features can be used for both purposes.

Due to the ambiguities discussed above, simplifying assumptions must be made about the correspondence maps. A good candidate for such an assumption is *neighborhood preservation*. Consequently, algorithms for correspondence finding usually optimize a combined objective function which favors similarity between local features and smoothness of the correspondence map.

One implementation of this procedure is *Elastic Graph Matching (EGM)*

(Lades et al., 1993), where stored models are represented as graphs vertex-labeled with vectors of local Gabor responses and edge-labeled with a distance constraint. The correspondence problem can be solved by optimizing the similarity between model graph and a (topologically identical) graph in the image in terms of similarity of both edge and vertex labels. This is a high-dimensional optimization problem, which is usually simplified by applying a hierarchy of possible graph transformations. It starts with pure translation, later adds scale changes, and finally local displacements. In the first steps, Gabor amplitudes are used exclusively, which leads to smooth similarity landscapes and allows for separating the different steps.

An alternative method, which makes use of the pyramidal form of Gabor wavelet transform is *Gabor Pyramid Matching* (Würtz, 1997). It starts with standard template matching of the Gabor amplitudes on a sparse grid and low spatial frequency and refines the results using higher spatial frequencies. Thus, neighborhood preservation is not explicitly coded into an objective function but inherited from the undistorted matching on low frequencies. Very precise correspondences can be obtained by subsequent subgrid estimation using the Gabor phases. This method allows for much better background suppression, because the need of knowing local features for each feature point on all scales is eliminated.

Memory organization

The importance of memory organization is due to the computational expense of the inevitable correspondence estimation, which should not be carried out separately on all stored models. Consequently, it is necessary to evaluate correspondences *between* the stored models. Adding this idea to EGM results in a so-called *bunch graph* (Wiskott et al., 1997). In that data structure, each vertex is labeled with one local feature vector from each person in the database and care has to be taken during creation of the bunch graph that these feature vector are indeed taken from corresponding points. In addition to different matching schemes, bunch graphs can be used in two major modes. In one mode, it is assumed that the person to be recognized is indeed in the bunch graph and selected according to similarity. Alternatively, the feature most similar to the given image can be selected for each vertex separately, leading to a composition of the face image in terms of local features of all persons in the bunch graph. Moreover, the vertices can carry additional

information like sex, beardedness, or a genetic disease of the person they belong to. By majority voting a decision about that feature for completely unknown persons can be made.

Eigenfaces (Turk and Pentland, 1991) are another technically successful approach to face recognition. Grey-value images of faces are prealigned by an optical flow method and then subjected to [PRINCIPAL COMPONENT ANALYSIS], which can be interpreted as a neuronal method. It turns out that a few components are sufficient to recognize identity. Recognition proceeds by projecting the image to be classified onto these components and applying a classifier to the resulting low-dimensional vector. Calculating the PC representation from a database of persons is rather time consuming but projection and classification is very fast. This shows that the major strength of the eigenface method lies in very efficient memory organization.

Neuronal models

On the technical side, there is quite a variety of neural network models applied to the problem of face recognition. They usually start from well-aligned faces with little variation. See (Gong et al., 2000) for a good discussion of the application of neural classifiers and an excellent treatment of technical approaches to face recognition.

It is currently not known if there is neuronal machinery in the brain to explicitly estimate correspondences. However, the [DYNAMIC LINK ARCHITECTURE] can be used to solve the correspondence problem as follows (Lades et al., 1993). Two layers of neurons that represent the image space in model and image, respectively, are fully interconnected by dynamic links. They have an internal wiring that supports moving localized blobs of activity. The development of links is supported by feature similarity and synchronous activation of the connected neurons. The link dynamics then converge to a correspondence mapping. It has been extended by a competition between a multitude of model layers to a full-blown neural face recognition system. This system is sped up by a coarse-to-fine strategy working on the Gabor pyramid. The speedup is due to the possible parallelism between all refinement steps. That system also shows good background invariance, because model- and image-representation are the same as for pyramid matching.

Face finding

Having found a correct correspondence map from a stored model into an image in principle implies that segmentation has also been solved. However, applying correspondence-based techniques like bunch graph matching to arbitrary images yields plenty of misclassifications — depending on the parameters, either many faces go undetected or many non-faces pass as faces.

It seems very difficult to encode the notion of a general face into a program or data structure. Therefore, for *finding* faces in images or video sequences neural net classifiers [APPROPRIATE HANDBOOK CROSSREF HERE]are widely used. Typically, a whole set of segmented and roughly normalized face images is used to train a network. Then the network is applied to all points of an image to provide a face/nonface decision. A good review of the facefinding literature can be found in (Hjelmås and Low, 2000).

Discussion

The notion of a correspondence map is a very general model of the variation of appearance of a face. Its estimation is computationally intensive, and the currently known neuronal models that can implement it cannot account for the rapidity of human recognition, even if run on highly parallel hardware like real neurons. The advantage of correspondence maps lies in the fact that much more information, like the actual position, pose and expression can be determined from them.

The quality of technical face recognition systems is difficult to judge. On small datasets (below 100 individuals) even naïve template matching may yield respectable recognition rates. The use of standard databases, inevitable for achieving fair comparisons, brings about the danger of overadapting the classifiers to the data. To prevent this, the Army Research Lab has set up a standard comparison procedure called the FERET test (Philips et al., 2000), where the major part of 14126 images of 1199 individuals is withheld for independent testing.

On this database, 8 competitors underwent a test with the additional information about the (hand-labeled) eye position. Only two competitors, a bunch-graph based system and an eigenface-based system took the realistic test on the images without any extra information. Both systems performed equally well on the dataset with given eye-coordinates, and the bunch-graph

based system clearly won on the more difficult examples without additional information (Philips et al., 2000). These results underpin the necessity of very good correspondence estimation for successful face recognition.

Roadmap: II.6 Vision

Background: Gabor wavelets

Related Reading: Gabor wavelets; invariant object recognition; dynamic link architecture; emotions; Face Recognition, Psychology, and Connectionism.

References

⋆ Chellappa, R., Wilson, C. L., and Sirohey, S. (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3(1):1–8.

Gauthier, I., Behrmann, M., and Tarr, M. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11(4):349–370.

⋆ Gong, S., McKenna, S. J., and Psarrou, A. (2000). *Dynamic vision*. Imperial College Press, London, England.

Grudin, M. A. (2000). On internal representations in face recognition systems. *Pattern Recognition*, 33(7):1161–1177.

⋆ Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). Distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233.

Hjelmås, E. and Low, B. K. (2000). Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274.

⋆ Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

Moscovitch, M., Winocur, G., and Behrmann, M. (1997). What is special about face recognition?: Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *J. Cognitive Neuroscience*, 9(5):555–604.

Penev, P. S. and Atick, J. J. (1996). Local feature analysis: a general statistical theory for object representation. *Network*, 7(3):477–500.

Philips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779.

⋆ Würtz, R. P. (1997). Object recognition robust under translations, deformations and changes in background. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):769–775.

Young, A. (1998). *Face and mind*. Oxford University Press, Oxford, England.