3 Models of face processing

In this chapter, specific attributes of faces and their basic parts, i.e. features and configurations will be discussed. The aim is to sensitize the reader to the specific properties of a face and its probable processing. For example, the salience of facial features differs significantly from one feature to the other with clear consequences for the recognition performance; especially the outer features differ strongly from the inner features. However, there are also important face-specific effects that must be considered. For instance, inverted faces are processed differently to upright faces, although the physical *Gestalt* does not change thereby. Similar tremendous changes come with the photographic negation of a face picture. A very important phenomenon with faces is the so-called *Thatcher illusion*. The perceptual effect of this illusion, which seems ideal for testing specific face processing hypotheses, will be used as a main paradigm for the experiments described in the empirical part of this work (see section 4.4 et seqq.).
Lastly, different processing assumptions will be introduced. Therefore, processing models will be explained in general and *face processing* will be discussed in particular.

3.1 Different kinds of information

A major deficiency in most accounts of face processing is their failure to spell out the *perceptual primitives* that form the basis of our representations of faces (e.g., Bruce, Burton, & Craw, 1992), and the underlying processing to recognize and combine these together to a coherent face *Gestalt*. We know that the human face is a highly meaningful stimulus that provides us with a great variety of information for adaptive interaction with people. Our ability to recognize faces is remarkably accurate, fast and long lasting (see for details the paragraph above). Many parts of the magic of face recognition have been identified by researchers, but the question as to which concrete steps have to be managed on the way to 'see a face', has mostly been excluded from investigation. Therefore, we will first consider the essentials of face recognition models in terms of *representation*, and discuss specific models of face *processing* later.

What kind of information is in a face?

There is quite a lot of information placed in a human face. From a physical point of view, the most informative dimensions in a face are featural aspects (e.g., the inner and outer facial parts as eyes, nose, mouth or chin and ears), the texture, the color[10] and last but not least, the

[10] In many studies, color and texture are defined as *local* featural aspects, as well. Color and texture information have been found to be specifically relevant for race and especially gender categorization (e.g. Hill, Bruce, & Akamatsu, 1995). In contrast to this, in the present work, 'features' are only specified as the inner components of a face, especially the eyes, nose and mouth.

relations/configurations between different features in the face (cf. Brooks, Don, Lewis, & Ragan, 1997). The main focus of the present work lies on the relation between configural and featural (componential) aspects of a face and the different importance between outer and inner features as well as the difference between upper and lower facial parts.
The first class, the term *component* or *feature* (or piecemeal, featural, componential) information, has commonly been referred to facial elements which are perceived as distinct parts of the whole, such as the eyes, mouth, nose, or chin (Carey & Diamond, 1977; Sergent, 1984b). In practice, the manipulation of component information has been achieved by replacing the components with different ones (Carey & Diamond, 1977; Sergent, 1984b) or by altering their color and shape attributes (Searcy & Bartlett, 1996; Leder & Bruce, 2000b).
According to Bruce (1988), the term *configural* information refers to the "spatial interrelationship of facial features" (p.38). In practice, the alteration of configural (or configurational, spatial-relational) information has been achieved by altering the distance between features (e.g., the inter-eye distance, Sergent, 1984b) or by rotating them (e.g., turning the eyes and mouth upside-down within the facial context, Bartlett & Searcy, 1993; Thompson, 1980). Configural information about position of features was analyzed in more detail by Diamond and Carey (1986). They used the term *first-order* relational information for the basic arrangement of the parts, whereas *second-order* relational information refers to specific metric relations between features.
The face as a whole can also be described as a template in which parts and spatial relations are contained only implicitly. The term *holistic* has been used for this kind of information (Tanaka & Farah, 1993; Farah, Tanaka, & Drain, 1995).

Prototype model

> "After a few years of traveling I meet a friend, and my first idea is 'How old he looks!' That does not at all mean that he looks particularly old on an absolute scale. […] Nor does it mean that I reproduce an image of him, as I knew him before, and now compare the two" (Köhler, 1935, p.271).

According to Vicki Bruce, the central problem in face identification is how we build stable representations from exemplars that vary, both rigidly and non-rigidly, from instant to instant and from encounter to encounter (Bruce, 1994). A number of authors (Goldstein & Chance, 1980; Valentine & Bruce, 1986a, 1986c) have suggested that faces may be encoded by reference to a generalized prototype or schema, that emerges as a result of a lifetime's experience with faces. Such a prototype is unlikely to be modified in the course of a single experiment with faces (Bruce et al., 1991; for an opposite view see Carbon & Leder, subm.), but it is a relatively stable representation, which plays the role of a basis. Beginning with this basis, new incoming information about already stored faces are handled as deviations from the prototype, which is a more powerful and economic kind of storage that is also used in computer systems for packing data and generating prototypes (Leopold et al., 2001; Basri, 1996; for non-face specific material see Minsky, 1975). Our internal representation of the outside world must be very sparse, because otherwise we would be confronted with an information overload, and the visual world would be a confusing collection of innumerous details, which would make it hard to experience a sense of continuity in our environment (Simons & Levin, 1997). As demonstrated recently with the paradigm of *change blindness*, it seems that our short-termed representations contain only what the observer is currently processing (O'Regan, 1999; Noë, 2002)[11]. Paradoxically, change (or inattentional) blindness suggests that we may be using the world as its own representation (O'Regan, 1992), thus the perceptual reality is constructed, and memory is the tool we use for that process. Only important objects become a part of a

[11] Recently, a critical opposition to this sight was presented by Cohen (2002).

person's visual representation of a scene, supporting the notion of a schematic and abstract representation of visual scenes that avoids unnecessary details (Intraub, 1997; O'Regan, 1999). What is represented in memory is not every detail, but rather the essentials of a scene.
However, what information does such a prototype hold? What are these 'essentials' that enable us to execute face processing tasks so well? In earlier days, by superimposing photographs of members of the same family, Galton (1879) was able to create a prototypical face with which each member had a 'family likeness', independent of the specific shape of single components. The so-called *prototype effect* reveals a similar effect: there is a tendency to respond to the central value of a series of varying exemplars, even when this central value or prototype has not been experienced. (Cabeza, Bruce, Kato, & Oda, 1999)[12]. Thus, it seems that especially configural information is essential for face prototypes. As all faces share a common spatial arrangement of features, variation arises from individual spatial deviations from a facial prototype (Diamond & Carey, 1986). Therefore, it seems very important to represent these fine differences in a prototype. Having such a representation with prototypes, an identification decision can be made on the basis of the relative goodness-of-match of the stimulus information with the relevant prototype descriptions (Solso & McCarthy, 1981).

What is a prototypical face?

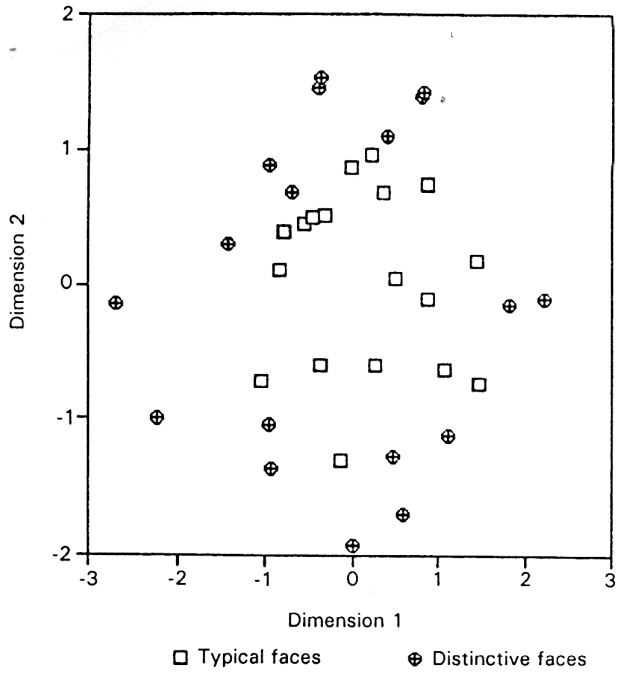
According to Klatzky (1980) a prototype in general is the most perfect or the ideal kind of a subject or object, stored in the long-term memory. It is rather abstract and pretends to be an averaged entity of a class or a central tendency. Other researchers found no evidence of a role for an abstracted prototype (Valentine, 1999).
A prototypical face is thought as being generated from a succession of all instances, which were seen as different pictures of one face. Like Baron's (1981) computer model of face recognition, the first prototype-like face will be encountered by the first exemplar of a particular face and will be stored as a discrete but not yet as a very stable pattern. When another exemplar is processed, it will be recognized as another picture of this person because of its invariant features (idiosyncratic features, configurations, textures, etc.). If it is identical with the first stored instance, then there is no need to memorize a new one. However, if it resembles it sufficiently to be categorized in the same way, yet deviates in some detected way from the instance already stored, then the original record may be replaced with an average of the two (Bruce et al., 1991). Recently, it could be demonstrated that this is a too simple assumption. Experimental data demonstrated that the latest incoming information concerning an already existing highly familiar prototype is weighted much more than older information is (Carbon & Leder, subm.). However, if the next exemplar of an already stored prototype is too dissimilar to be categorized in exactly the same way, it may be separately recorded. Thus, a set of more variant instances will be more likely to result in the storage of a set of exemplars (Bruce et al., 1991). In accord with the prototypical effect, infants do not prefer their mother's face but the prototype face. It is postulated that the prototype face is constructed as the sum of infants' visual experience, and in a natural situation the mother's face mostly resembles this prototype face (Hess & Slaughter, 1986). For elderly people it seems that this impressive adaptive representational system is not so flexible any more. There are significant age-related differences in the use and the generating of such prototypes yielding a higher false-alarm rate in the recognition of faces for older people performance (see also Inn, Walden, & Solso, 1993).
An interesting finding about prototypes comes from the attractiveness research. Attractive faces are close to the mathematical average of the population, which suggests that newborns prefer prototypical instances of faces (Langlois & Roggman, 1990). Therefore, a prototypical face is also a face that is centered to high attractiveness.

[12] For a critique see Schacter, Verfaellie, and Anes (1997).

Face space

The theory of prototypical faces is based on one key assumption, called the *prototypicality hypothesis*. According to this hypothesis, faces are organized in a *face space* with the most prototypical face in the center. Additionally, it is assumed that the more a face is dissimilar to this prototype, the more distant it is from this *center* or *average point* (Johnston, 1997). Rosch and colleagues (Rosch, Simpson, & Miller, 1976; Rosch & Mervis, 1975) have demonstrated that rated prototypicality is highly correlated with interitem similarity: "members of a category come to be viewed as prototypical of the category as a whole in proportion to the extent to which they bear a family resemblance to [...] other members of the category" (Rosch & Mervis, 1975, p.575). Therefore, the *psychological face space* can be measured by asking subjects to make pairwise similarity ratings between all pairs of faces in an experiment and submitting the results to a multi-dimensional scaling (MDS) procedure (e.g., Henss, 1994; Hojo, 1988; Sergent, 1984b; Valentine, 1999; Busey, 1998).
Empirical evidence shows that faces that are more similar to other faces are less easy to recognize (Light, Kayra-Stuart, & Hollander, 1979). Other researchers demonstrated that even reaction time (RT) predictions are possible with a given face space (Takane & Sergent, 1983). The iterative goal of an MDS is to find a configuration of points in some multidimensional space such that the interpoint distances are monotonically related to the experimentally obtained (dis-)similarities (Kruskal & Wish, 1978; Steyvers & Busey, 2000). As a consequence of the key assumption, the structure of the face space is the following: typical faces have to be located near the center, and distinctive faces having idiosyncratic features or configurations must be most distant to the prototype. This was empirically validated by Valentine (1991) and others (Johnston, Milne, Williams, & Hosie, 1997; Johnston, 1997). See an illustration of empirical data from Johnston et al. (1997) in Figure 3-1. A similar investigation of face space will be done in Pre-Study 1a and Pre-Study 2 in the empirical part of the present work.

Figure 3-1: Graph of two-dimensional plot of face-space from Johnston et al. (1997, Figure 2). The distinctive faces are located in the periphery of the face space (symbolized as crosses with circles). In contrast to this, the typical faces will be found in the central area of the face space (symbolized as quadrates).



Nevertheless, which dimensions are represented in the memory for faces or the *face space*? With the help of data compression techniques like the principal components analysis (PCA), these dimensions can be extracted mathematically (Hancock, 2000). Besides this, PCA delivers a radical data compression (Burton, Bruce, & Hancock, 1999). As a matter of fact, faces could be represented in very few dimensions (Kirby & Sirovich, 1990; Valentin, Abdi, & O'Toole, 1998), but the PCA could not extract all important dimensions. A particular problem is that PCA is not capable of extracting local feature-like structures in objects, but Penev and Atick (1996) successfully showed how to construct from the global PCA modes a local topographic representation of objects in terms of local features in a *Local Feature Analysis*.

A structural problem of the face space account is the assumption of *one* singular face space. For example, recognizability and gender classifiability of faces are commonly identified as being independent. Therefore, these two measures cannot lie in the *same* face space (O'Toole et al., 1998). This inference can also be drawn from other experimental data, where different tasks generate seemingly different face spaces (Leder & Carbon, subm.-b). Moreover, other findings challenge the proposition that the feature space ever becomes fixed (Williams et al., 1998). An expert's feature space may become reorganized in response to environmental pressures to perform a categorization task more efficiently (evidences for this high adaptivity Lowe, 1985; Tsotsos, 1990; Tsotsos, 1995). Therefore, the face space model can only give us a snapshot of the *current* representational structure of the *current* task.

However, despite all this theoretical criticism, there is also very much support for the idea of a psychological face space. For example, the theoretically assumed higher exemplar density of other-race faces indeed makes recognition more error-prone in the same way that recognition of typical faces is more error-prone than recognition of distinctive faces (Valentine, Chiroro, & Dixon, 1995). Ellis, Deregowski, and Shepherd (1975) found that different facial features are used to describe black and white faces. Shepherd and Deregowski (1981) found that when three faces of the same race are presented simultaneously, the facial features used to judge similarity among black faces differed from the facial features used to judge similarity among white faces. Compared with own-race faces, other-race faces seem to be clustered more densely in the multidimensional space. Generally, it is found that the classification accuracy declines with distance from the prototype (see Light et al., 1979). There is a similar effect concerning caricatures. It is commonsense that caricatures hold the characteristics of a face in a more distinctive way and therefore can be recognized in a more efficient, accurate and even faster way (Benson & Perrett, 1994; Knappmeyer, Cheng, & Bülthoff, 2002; Mauro & Kubovy, 1992). The *caricature effect* is a robust finding, which has been demonstrated employing different tasks and different stimuli, such as line-drawings or photo-realistic morphs.

Caricature line drawings include stable features and emphasize distinctive ones, despite the impoverished and distorted nature of the information in caricatures (Benson & Perrett, 1991). As such, caricatures are closer to schematic memory representations than are photographs (Tversky, 1985). If caricatures are closer to memory representations than photographs, they should lead to better memory and faster retrieval. In a study of Benson and Perrett (1994), 42% of the caricatures were considered to be best likeness of famous individuals, and more distinctive faces required less caricaturing (Exp.2). As a consequence, caricatures should be found in the periphery of the face space with high dissimilarity ratings in respect to other faces. Indeed, caricatured drawings provide improved access to memories of famous faces, which gives strong support to models of human-face memory and processing based on norm-based coding (Benson & Perrett, 1994; Rhodes, 1995; Rhodes, Byatt, Tremewan, & Kennedy, 1996). The range of a face from its 'anticaricature' (a face with all features morphed towards the prototypical face) to its caricature seems to be a direct continuum, starting in the centerpoint of the face space (Leopold et al., 2001).

There are many rival models of how the face space is organized. There is the rivalry of *viewed-based* or *individual-based* accounts (Newell, Chiroro, & Valentine, 1999) and that of

the norm-/prototype-based coding model and the purely exemplar-based model (Valentine, 1991). The debate between exemplar- and prototype-based learning has been intense in the literature on concept formation (e.g., Smith & Medin, 1981). It was argued that models based on a combination of exemplars and *abstracted* prototypes were likely to be most successful on the long run. Homa, Sterling, and Trepel (1981) presented evidence for such a 'mixed' model for the learning of abstract patterns, whose shapes vary quantitatively (cf. Bruce et al., 1991). Other researchers, differing between *template-based* and *feature-based* accounts, have also found that chimerical models are most successful for automatic face recognition systems (Brunelli & Poggio, 1993). What is common to most automated solutions that have yet been implemented is that they integrate more information than pure configural or pure featural ones (see Beymer & Poggio, 1996; Blanz & Vetter, 1999; Wenger & Townsend, 1999). A different approach was investigated by Lades et al. (1993), who developed a face recognition algorithm based on the *dynamic link architecture*, that have only modest capabilities to generalize over large rotations in depth (Biederman & Kalocsai, 1998). This algorithm combines alignment strategies with identification processing (see also Wiskott, Fellous, Krüger, & von der Malsburg, 1999).
Additionally there is a controversy about the metric and the dimensionality of the space. In Valentine (1999) and Valentine et al. (1995), it was assumed that the metric is most probably Euclidean. The dimensionality debate concerns the question what features underlie face processing, and it could give us insight into the kinds of features that may prove most efficacious for this task (Cottrell, Dailey, Padgett, & Adolphs, 2001). It is extremely advantageous to further analyze the structure of a 2D-face space, because we can better imagine such a simple space than a space with higher dimensions. For a full discussion on this point see the detailed literature (Craw, 1995; Wenger & Townsend, 1999).

Local and configural processing

There are some hints that configural and featural aspects of a face are dissociable and are *not based* on the same cognitive processes. One strong piece of evidence comes from faces presented upside down, so-called *inverted* faces. Inverting isolated facial features had little effect on the recognition rate, whereas inverting whole faces resulted in a significant decrease of the recognition performance (Tanaka & Farah, 1993). This could also be demonstrated by configural vs. featural-altered faces. Again, the recognition rate dropped for configurally changed faces ('configural faces'), but was stable for 'featural faces', when inverted faces were compared with upright ones (Leder & Bruce, 2000b). This implies that there are two different coding systems: one which is unaffected by inversion (featural), and another which is disrupted by inversion (configural).
Moreover, face inversion eliminated the *prototype effect* for configural prototypes but not for featural prototypes (Cabeza & Kato, 2000). Participants who viewed the faces inversely, selected more of the featural faces than of the configural faces, but there was no difference with upright faces. This again demonstrates the dissociation between the two manipulation classes.
The results of clinical research studies confirm the dissociation hypothesis, too. For example, autistic patients showed significantly *more* activation than the control group in the right inferior temporal gyri, and *less* activation in the right fusiform gyrus during a *face* recognition—but not during an *object* recognition test (Schultz et al., 2000). It is likely that individuals with an autism spectrum disorder process faces in a different manner than normal subjects, relying more on featural than on configural analyses (Hobson, Ouston, & Lee, 1988). Contrary to this, the pattern of the brain activities of neurologically normal subjects seems to be specialized for configural processing (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher et al., 1997). Such differences in perceptual style among persons with autism would be consistent with people lacking face expertise, or young children and old people (cf. section 2.2.2).

Confoundation of local and configural information

Many experiments explicitly differentiate between *configural* and *local* face manipulations. The commonly used method for configural changes is to shift features within the facial context, and for featural changes to exchange single face components (e.g., Haig, 1984; Haig, 1986). Recently, a new technique of McKone, Martini, and Nakayama (2001) operating with noise patterns also tries to isolate configural from featural alterations.
However, as Sergent (1984a) has already pointed out, feature manipulations inevitably affect the configuration of a face (cf. Bruce et al., 1991). Lewis and Johnston (1997) argued in accord with this, giving a convincing example: "The presence or absence of a big nose can be featural, but if one considers the nose to be big only in relation to a small face then configural coding is required" (p.225). The hypothesis of local and configural confoundation could lately be demonstrated empirically: Leder and Bruce (2000b) created pure featural versions by different coloring of the features, entire configural versions by shifting the same features among all faces, and a mixed configural plus featural manipulation by exchanging the features between the different faces. The last condition corresponds to the versions commonly accepted as *featural*, in which configural and featural manipulations are confounded. The inversion effect, which can be seen as an indicator of the amount of ongoing configural recognition processes (Le Grand et al., 2001), was the largest for the entire configural versions and nearly non-existent for the pure featural version, with a medium effect of the 'confounded' version (Leder & Bruce, 2000b; Leder & Carbon, in prep.).

3.1.1 Feature-based

> "On 15 November of our evolutionary year, there emerged a creature called Pikaia. [...] it had a hole at one end. [...]. The simple opening was a primitive mouth [...]. But eyes had not developed yet. The face—as we know it—began with a mouth. [...] the Conodont was the first creature to have two 'eyes'. [...Then] they developed primitive ears, [...] then came a nose" (Bates & Cleese, 2001, p.14 et seqq.).

As we see in this short 'evolutionary story', the individual features of a typical face were developed in different time periods under diverse selectional instances of pressures from the outside world. For our social life, the recognition of these features is of varying importance, too (see Parks, Coss, & Coss, 1985). For example, the ears and the nose have no possibility to give us social cues, because their movemental repertoire is very limited, apart from some artistic performances in the circus. Conversely, the muscles around the eyes reveal the secret of true or false laughing (Ekman, Friesen, & Ellsworth, 1972), give us a hint about the emotional involvement and the state of health of a person. The same is true of the mouth, also a highly informative facial feature. In an archaic society or natural culture, these cues were essential for the survival of the fittest. The fast and accurate evaluation of the mood, physical ability and aggressiveness of the enemy played an important role. Generally speaking, the things to which we attend are said to interest us: "Our interest in them is supposed to be the cause of our attending" (James, 1905, p.416).
Therefore it seems plausible, as Rakover (1998) and Rakover and Teucher (1997) suggested, that featural information is the most important information for face recognition, with an even higher informative value than configurations. Moreover, because of their highly emotional and social content, they should be analyzed very quickly.
The Irving Biederman group tested this working hypothesis by exposing participants to sequentially very briefly presented pairs of line drawings of simple objects, that were composed of two parts (t=100 ms, followed by a 100 ms mask) (Cooper & Biederman, 1993). In a task, in which subjects had to judge whether the two images had the same or a different name, it was shown that NAPs (Non-accidental properties; that is a form or featural changing) had a greater salience over metric properties (see also Biederman et al., 1999).

Many other findings suggest that piecemeal-feature or part-based processing might also play a role in face perception and memory, in some cases a strong or even dominant role (Bartlett, Searcy, & Abdi, in press). It would not be surprising if processing started at the level of individual elements (Parks et al., 1985). Features, after all, can easily be recognized as 'wholes' themselves, even in isolation. And if embedded in the global properties of a face, even infants are able to do so (Meltzoff & Moore, 1977).

Do features have interactive or independent characteristics?

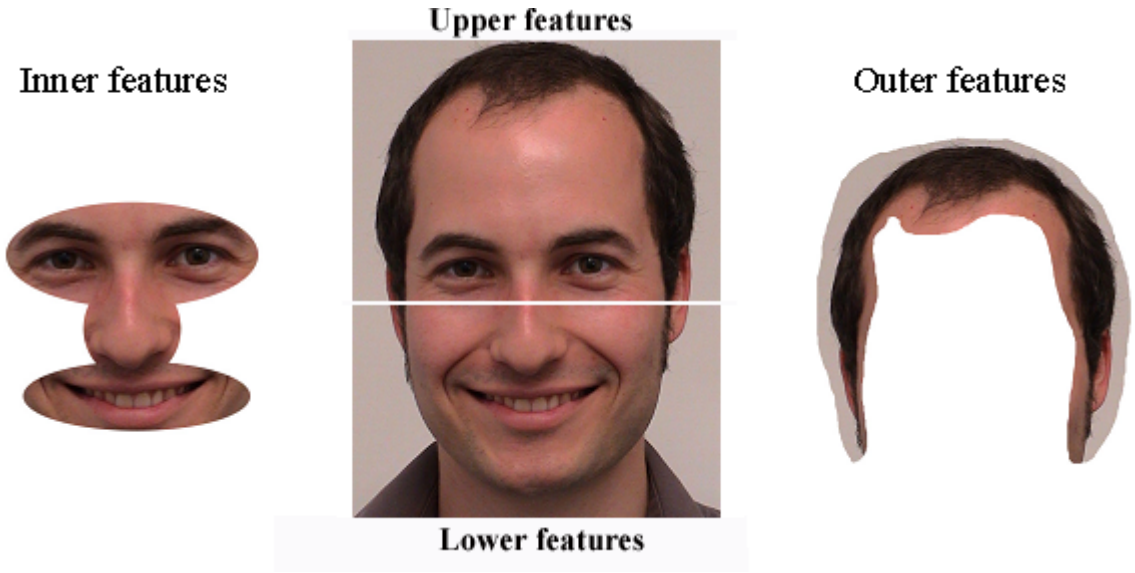
Young, Hellawell, and Hay (1987) split pictures of famous faces into separate top and bottom halves which they re-paired to form new, composite faces. They found that subjects were very impaired at naming the top halves of the composite faces when these were closely aligned with the 'wrong' bottom halves, compared with when the two halves were misaligned (see also Hole, 1994). Somehow, the new configuration formed by combining upper features with the wrong lower features overrides subjects' abilities to recognize the features independently (Bruce et al., 1991). But this was only found when such composite faces were presented upright: Sergent (1984b) found parallel results for upright faces with significant interaction between different feature areas of a face, but the interaction between face features was removed by inversion (cf. for separate processing of emotional features in inverted faces in Treisman, Cavanagh, Fischer, Ramachandran, & Von der Heydt, 1990). Furthermore, there are also evidences for interactive processing from face expression research. Different facial areas interact to determine a particular emotion (Ekman et al., 1972), for example, the same movement of the brow may convey different emotions depending on the movement of the mouth (McKelvie, 1973). Additionally, Sergent (1984b) interpreted RT data in favor of an interactive account for upright face features. She found a decrease in RT with an increase in the number of differences between faces, which could be a result of self-terminating analytic processing, but was also interpreted as being consistent with a configural interactive mode of processing.
In contrast to this, Tversky and Krantz (1969), using extremely schematic faces, found opposite results. Their participants had to rate the dissimilarities of single facial features (eyes, mouth, and facial shape). The separate dissimilarity scales were found to be independent from each other. Both studies, Sergent (1984b) and Tversky and Krantz (1969), used *Identikit*, and highly schematic faces, respectively. However, Leder (1996) uncovered schematic faces as being problematic stimuli in face processing experiments, because they might differ qualitatively from natural faces in respect to the underlying processing. Therefore, Macho and Leder (1998) re-tested the interactive processing hypothesis, now with realistic faces. Evaluating their similarity rating data by the *logit* model, they failed to find interactive feature processing, which they considered as a possible explanation for holistic processing. Similar results were found by Rakover and Teucher (1997), who assumed that recognition of a whole face is a function of separate recognition of one of the five features, used. In addition to this, different feature dimensions like color, size, orientation and spatial frequency were also found to be processed quasi-independently (Uttal, 1981; Bruce, Green, & Georgeson, 1996).

Features in the upper vs. lower face area

Because of the different nature and informative content of the single features, it makes sense that we should have developed an analyzing technique that optimizes the 'investigating look' at the *regions of interest* (Kanwisher et al., 1997), those areas, where most information for the current task is available or supposed to be available. Besides this, humans generally show a preference for novelty. The greater the neural activity the greater the endorphin release and thus the more pleasurable the experience (Lewis et al., 1981). Moreover, the more pleasurable it is, the more informative it seems to be. However, where in a face are such regions of interest found?

Many studies seem to indicate that there are differences of relative importance between certain facial features for the recognition of faces. The studies are consistent in showing upper-facial features as playing a dominant role in the recognition of faces (e.g., Lacroce, Brosgole, & Stanford, 1993; Fraser & Parker, 1986). An illustration of upper vs. lower features as well as outer vs. inner features is shown in Figure 3-2.

Figure 3-2: Illustration of upper and lower features as well as outer and inner features of a face.



Upper-facial features are usually defined as hair and eyes, while lower-facial features are defined as the chin, nose, and mouth (Reynolds & Pezdek, 1992).There seems to be a hierarchy that explains the role of each feature in recognition. This ranking system was studied and explained by Haig (1984). Haig performed a series of studies that were designed to measure the relative importance of different facial features. He found the head outline as the dominant recognition feature. Next in importance is the eye/eyebrow combination, followed by the mouth, and then the nose. The relative importance of the different features in recognition was determined by changing one feature between study and test faces in a recognition task. Importance of a particular feature was ascertained by calculating the percentage of correct responses to the test faces. In this sense, a higher percentage of correct responses indicated a greater importance of the feature (Haig, 1984). People's recognition rate dropped substantially when the shape of the head outline was changed from that in the study face, followed by changing the eyes region. This seems to indicate that the head outline gives people more information to make a discrimination decision than do the eyes. However, the eyes region, representing the upper part of the face in this experiment, was still much more prominent for correct recognition than the lower parts of the face. This is in full accord with early studies by Smith and Nielsen (1970), who found that their participants used a top-to-bottom scanning strategy, paying particular attention to the upper face features (see also Goldstein & Mackenberg, 1966; McKelvie, 1976; Davies, Ellis, & Shepherd, 1977).

Inner vs. outer features

In other studies, it was found that internal[13] facial features were more important in recognizing famous (familiar) faces, while unfamiliar faces were recognized using either internal or external features (Ellis, Shepherd, & Davies, 1979). This different pattern of results dependent on features raises important issues about how the features interact and about the role that ho-

[13] The internal features the present work will focus on are the following: eyes, nose, and mouth.

listic recognition or the recognition of the outer contours plays in face recognition. However, it seems clear that the importance of these categories of features is dependent on the familiarity of the given faces, a factor that was not included in the Haig (1984) study cited above, where the outer face context was identified as most important for face recognition in general.

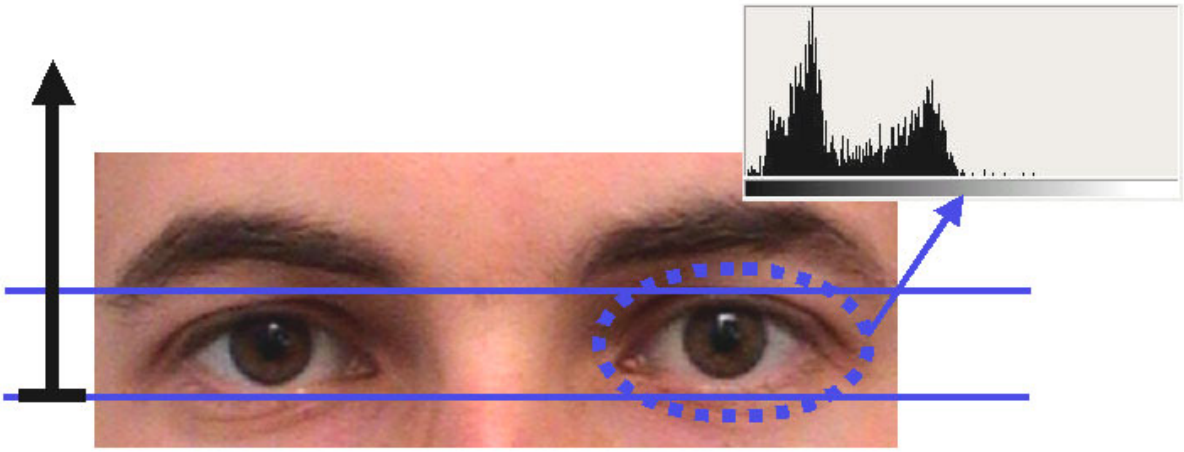
Inner features

Instead of focusing all objects simultaneously, processing is often limited to one object in a certain part of space at a time, e.g., through focusing attention on a *region of interest* in the visual field, which is then routed to higher visual areas (Riesenhuber & Poggio, 1999a). With respect to faces, such a high interest area is the field of the eyes, because of its physically and socially discriminative value (Brunelli & Poggio, 1993). But also the mouth is very informative, because of its high physical contrast and its dynamic and socially important nature (Ellis, 1975; Scassellati, 1998). Eyes may constitute a special stimulus in at least two senses. First the eye's structure is such that it provides us with a particularly powerful signal to the direction of another person's gaze, and second, we may have evolved neural mechanisms devoted to gaze processing (Langton, Watt, & Bruce, 2000; Langton & Bruce, 2000). Other features seem to be not as important as the eyes and the mouth, for example the nose or the cheeks, which are relatively stable. The unequal contribution of the features to the recognition of a face could therefore concur with the view that an analytic mode of processing prevails in facial perception (Ellis, 1975; Davies et al., 1977).

Recently, electrophysiological studies in epileptic patients have found regions of the temporal cortex, that respond to socially salient parts of faces, such as eyes and moving mouth parts (Adolphs, 2001). This could also be demonstrated by observing the N170 component (see chapter 2). By testing children, the N170 had shorter latency and was much larger to eyes than faces, suggesting the early presence of an eye detector (Taylor et al., 2001). Similarly, in a *face-identity* decision task, the eye region was more heavily weighted than others, but only when the face was intact and upright (Tanaka & Farah, 1993). Thus, it was not a pure *eyes*-effect but rather an *eyes-in-a-face* effect. It was also found that the eyes and their outline were more salient than the nose and mouth in terms of both, speed of processing and error rate (Fraser, Craig, & Parker, 1990; Fraser & Parker, 1986). Rhodes (1988) proposed that faces are recognized sequentially in a top-down ordering from forehead to eyes, to nose, mouth and chin, and thus the eyes play an extraordinary role for the recognition process. This can be strikingly demonstrated by the frequently used method to anonymize persons by simply covering the eyes region with a black bar. The high importance and signal character of the eyes is also immanent in flora and fauna. For example, peacocks as well as butterflies present eyes as attracting and warning signals on their feathers and wings, respectively.

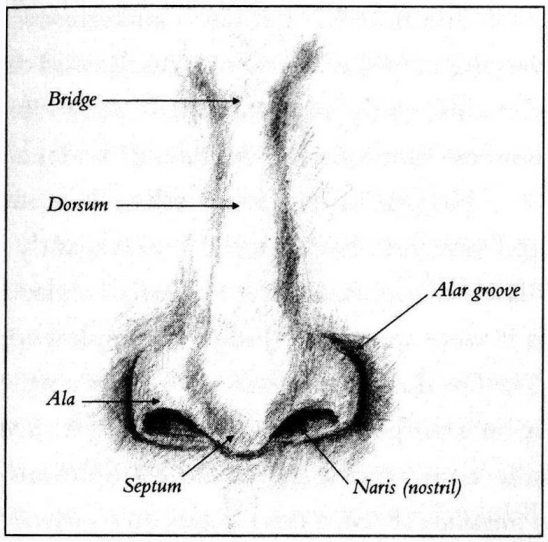
However, there is another very significant property of the eyes. Because of their high contrast and their symmetric positioning, they are ideal as an indicator for (a) the existence of a face and (b) the current alignment of this face. As Kosslyn (1973) pointed out, it is useful and very economical to navigate to the most informative point, because it is often also the most relevant one. As demonstrated in Figure 3-3, the eyes build an artificial horizon that facilitates the alignment of the face. This is possibly advantageous for recognition of the *whole* face, because without a valid alignment method, it is difficult to compare the visual pattern inside a well-known coordinate system and to quickly locate other less dominant features. This method is also commonly used by recognition machines (Scassellati, 1998; Huang, Liu, & Wechsler, 1998; Minut, Mahadevan, Henderson, & Dyer, 2000; Teh & Hinton, 2000; Brunelli & Poggio, 1993). The advantage of a simple symmetry detector is that it does not require the knowledge of the object's shape. Nevertheless, it needs valid anchor points like the eye-to-eye axis or the mouth-eyes triangle (Zhao et al., 2000; Fromherz, Stucki, & Bichsel, 1997; Intrator et al., 1995).

Figure 3-3: Demonstration of the high informative content of the eyes region. First, it is a valid indicator for the alignment of the whole face. As symbolized by the two parallel blue lines, the x-axis of the face can easily be found by the connected lines between the eyes. Second, the eyes are very salient because of their complex visual pattern and the strong contrasts, found in the inner eyes region. This is demonstrated by a luminance histogram of this region in the image-editing program Photoshop (Adobe Systems Incorporated, 1997).



The high apparentness of the eyes is mostly the result of the immense contrast spectrum of this region. With the help of a nearly white sclera and a completely black pupil, it is a highly significant cue. In Figure 3-3 you can also see a luminance histogram of the inner eyes region with the sclera, the iris and the pupil. It is obvious that there are two significant peaks in the histogram: the one in the darker luminance range represents the black pupil and the brown iris; the other one in the lighter luminance zone represents the white sclera. On the whole, the eyes tend to be darker than the surrounding face, which underlines the contrast relative to the whole face, and the outer world even more (Scassellati, 1998; Lowe, 1985; Baylis & Cale, 2001) and accentuates the figure-ground relationship (Hoffman & Singh, 1997).
Another possible indication for the outstanding prominence of the eyes can be found in the results that eyes were recognized independently of other facial features (Macho & Leder, 1998) and in the recognition rate data of many other studies. Recently, Leder and Carbon (subm.-d) showed in a face recognition test that part-based presentation of the eyes alone still yielded very high identification rates (.781), but dropped dramatically when only mouths (.470) or noses (.340) were presented. This result pertains to the hypothesis that not all facial features are equally important in frontal views of the face (cf. Macho & Leder, 1998; Bruce, 1988). However, the reason for this weak performance seems *not* to be a bad representational system for noses and mouths as such, because when these features were learned in isolation, much higher recognition rates succeeded. It seems more that there exist two different elaboration modes, one in a holistic style when faces are shown as whole faces, and a more part-based style when the faces are shown in their part-based versions (Leder & Carbon, subm.-d).
Despite the fact that a nose does not hold a wide contrast spectrum and does not operate as an important social and identifying cue, there are astonishingly sophisticated descriptions of this area (see Figure 3-4).

Figure 3-4: The nose (figure from McNeill, 1998, p.31).



The outline of a face

It is very important to categorize very early attended visual percepts as faces and as identities in order to save further calculations and to have recourse on the semantic network (Zhao et al., 2000). Therefore, coarse stimulus information seems to be ideal, because it is not necessary to identify any fine details but a simple global pattern. In accord with this, contour was the fastest feature to discriminate between identical-looking faces (Sergent, 1984b), that is, latencies to a contour difference were significantly shorter than latencies to either eye or internal configural difference. Similar results were found by Sergent (1984b) with a changed chin as an outer feature being faster recognized than internal features like eyes or mouth (cf. Bruce et al., 1991). It can also be demonstrated that outer and inner features do not only differ in salience and speed of recognition, but also seem to be of a different quality. For example, Leder and Bruce (1998) showed that effects of distinctiveness, based on manipulations of local or configural internal facial features, can be observed much more intensely when the outer features are omitted from the stimuli. Moreover, the recognition of faces which differed only with respect to selected internal features was very sensitive to orientation, and thus increased the face inversion effect (Leder & Bruce, 1998).

Relation between inner and outer features

Fraser and Parker (1988) found that facial fragments presented to observers in sequences were most correctly perceived when the outline was presented first or last. This emphasizes the special importance of the outline in pattern recognition and supports the existence of two processes in visual perception: a high-level parsing process, and an outline priority effect. Both processes are probably precognitive in nature (Fraser & Parker, 1988). This was replicated in another study, where the median reaction times for all correct responses to feature omission were calculated. The overall means across all trials indicated that responses were faster for the detection of the omission of the outline as opposed to the eyes, the nose and the mouth (Fraser & Parker, 1986).
However, there is additional dissociation between inner and outer features. In accord with the results listed above, in the case of a jumbled face the external pattern was again dominant when a face was *upright*. When *inverted*, the results were reversed and thus the internal pattern was dominant, then (Rakover, 1999). Moreover, the recognition of internal features was more influenced by inversion than the recognition of external ones, which, by the way, resulted in a low sensitivity for expressions in inverted faces (Phillips, 1979). The reason for

this effect lay probably in the high importance of configural information for internal features, whose processing is hindered by inversion (Leder & Bruce, 2000b). In contrast, the outline consists mainly of gross features without much configural subtlety and thus was not affected as much by orientation (Sergent, 1984b). Interestingly, when *only* internal features were available by omitting the surrounding context, effects of distinctiveness based on manipulations of local or configural internal facial features came out much stronger (Leder & Bruce, 1998).

The outer face line is not always predominant, as discussed above. For example, adults based recognition of *familiar faces* more on their inner features (eyes, nose, mouth, cheeks) whereas recognition of *unfamiliar faces* is based more on the outer features of the head (hair and oval face shape) (Ellis et al., 1979). Other researchers also demonstrated, using the *repetition priming* paradigm, that recognition rates for internal features were significantly higher than for external features (Brunas, Young, & Ellis, 1990). Thus, one cannot obligatorily speak of a *general* predominance of outer features of a face.

Salience[14]

A number of studies have shown that reaction times (RTs) to a stimulus decrease as the intensity of the stimulus increases (Massaro, 1989), so the nerve conduction velocity is not—as Helmholtz assumed—constant (about 30 m/s) (see for details Massaro, 1989, p.133). Therefore, *salience/distinctiveness* seems to be an evident variable for face processing, too (Shapiro & Penrod, 1986; Leder & Bruce, 1998).

A striking evidence for treating distinctive faces differently from typical ones is an evaluation study, in which people had to rate the age of normal and caricatured faces. Caricatures, pictures that are inherently more distinctive than their related original pictures, were judged older than the veridical faces (O'Toole, Vetter, Volz, & Salter, 1997). In other words, this study indicates that the concept of *distinctiveness* has a much greater and more complex impact than its pure salience.

However, as mentioned above, distinctiveness has also a significant impact on the recognition speed of a face and its features. Distinctive faces were recognized *faster* but were classified as faces more *slowly* than were typical faces (Valentine & Bruce, 1986a; cf. Steyvers & Busey, 2000). Parallel evidences stem from accuracy measurements. There were more hits and fewer false alarms for unusual faces than for typical faces (Light et al., 1979; Going & Read, 1974). Therefore, faces seem to be encoded by reference to a general face prototype. A salient face is located more distant from this prototype and therefore will 'pop out' as highly flashy (cf. section 3.1). However, in a classification task, the greater distance of a high salient face made it also more untypical-looking and therefore increased the RT (see Valentine, 1999). These findings for full faces can be transferred to facial features. There is a large body of literature suggesting that facial features are not equally distinctive and therefore not equally effective in assisting face recognition (Ellis, 1975; Davies et al., 1977; Bruce, Dench, & Burton, 1993; Rakover & Teucher, 1997).

Locating distinctive regions and features is an important ability of the visual apparatus. By detecting luminance changes, new or informative elements can be identified. Participants in a *multiple-element display* experiment were able to prioritize up to 15 elements simultaneously (Donk & Theeuwes, 2001). The visual routines identifying these *regions of salience* or *coherence* in the visual field work preattentively and in parallel (see Shepherd, Davies, & Ellis, 1981). These regions are then subjected to further analysis by focal spatial attention process-

[14] In the present work, no conceptual difference is made between the terms *salience* and *distinctive* or *unusual* (like in Bartlett, Hurry, & Thorley, 1984); for similar questions, other authors used the terms *atypical* or *unique* (Light et al., 1979), and *typical* or *average* for indicating their opposite. Nevertheless, in the empirical part of this work, *salience* will be used as an independent variable, whereas *distinctiveness* will be used as a dependent measure.

ing (Julesz, 1981b; Treisman, 1982). While humans perform visual detection and localization in parallel, discrimination has to be done serially (cf. section 3.2.3). Moreover, people do tend to attribute different weights to different facial dimensions and features (Shepherd et al., 1981; Bruce, 1988; Rakover & Cahlon, 1998). So it is highly probable that humans have specialized pre-formed *saliency maps* (Treisman & Gelade, 1980) of the face in mind, which, however, are flexible enough to react to untypical saliencies that are to be triggered when recognizing a face. This is one of the critical differences between human vision and less efficient machine vision algorithms. Our visual system can decrease its computational load by automatically selecting relevant locations, while artificial vision systems must generally explore the visual world in a more systematic fashion (VanRullen, in press). Therefore, visual saliency is a fundamental yet hard-to-define property of objects or locations in the visual world.

Does all salience lie in the stimulus itself?

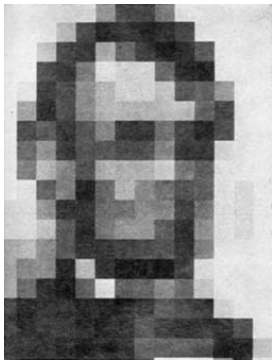
J.J. Gibson was concerned to show that the environment is a richer source of information than many perceptual psychologists had realized before (e.g., Gibson, 1950; cf. *theory of cognition* Neisser, 1976). In particular, he argued that the environment contains a variety of high-level invariants, such as ratios, proportions and salient areas, that do not alter under local changes in optical stimulation. In this sense, perceptual development and perceptual learning are seen as a process of distinguishing the features of a rich input, not of enriching the data of a bare and meaningless input (Gibson, 1964). Therefore, objects are specified by their inherent properties and do not consist of the subjective memories or the interpretation through the apparatus. So, the recognition and memorizing of an object is more a case of attention guiding than a compilation of association (Gibson, 1979). Unfortunately, Gibson failed to explain how we perceive invariances, because he did not explain how we derive representations of invariances from *fleeting sensory information*[15]. However, Gibson also showed that objects like faces have their own intrinsic qualities and saliencies, which are not explainable by the cognitive system alone. Moreover, the physical salience of a feature is not enough for it to be processed fast and accurately. As demonstrated by the *change blindness* paradigm, it is necessary that the attentional focus be guided to this feature, and the content of this feature must be very important. Otherwise, there is 'seeing without looking', or like O'Regan (1999) described it: "The eye may for example be fixating at the center of a circle, where there is nothing to be seen, in order to check that the circle is round rather than an ellipse" (p.85).

3.1.2 The holistic approach

Our capacity to identify a face as far as 40 meters away (Sergent, 1984b) suggests that the visual system must rely on characteristics other than single features, which are not dissolvable any more under such limited conditions (see also Lakshminarayanan et al., 1997). At such a distance, only the relationship among these features and the overall appearance may provide the cues necessary for identification.
Leon Harmon (1973) was able to demonstrate that the human being is indeed capable of identifying even a very sparse sight of famous faces, like that legendary one of Abraham Lincoln, which was presented in only 14 x 18 quantized squares (cf. Bachmann, 1991). See Figure 3-5 for an illustration of this effect.

[15] The *fleeting sensory information* is defined as all that we receive (Kitcher, 1990).

Figure 3-5: The former U.S. president Abraham Lincoln in a demonstration by Harmon (1973), where the original picture is degraded to a 14x18 pixels resolution. For people who know the visual appearance of Abraham Lincoln, it should still not be any problem to identify him.



At such a low level of resolution, no single part of the face conveys enough information for identification, and it is only the configuration or the *holistic something* resulting from the coarse quantization of the stimulus that allows any recognition. Of course, such a high face perception performance is only possible for face experts like adults (Schwarzer, 2000).
Within the holistic processing theory, it is hypothesized that face processing relies more on (automatic) holistic representation than does the processing of other stimuli (Tanaka & Farah, 1993; Farah, Tanaka et al., 1995; Farah et al., 1998)[16]. This idea is based on the idea of *Gestalt psychology* (Wertheimer, 1938)[17] that the whole is more than the sum of its parts, and translated into a template-like processing approach. According to this position, a face is seen as an organized, meaningful pattern, that is a more potent stimulus than an arbitrary assemblage of the same visual features (cf. Purcell & Stewart, 1988).
Recently, this position enjoyed its renaissance, thanks mainly to the work of the Tanaka and Farah research group. These researchers propose that faces, more than any other class of objects, are recognized on the basis of a representation in which parts are less explicitly represented than in other objects. Faces are proposed to be a prototypical class of this kind of holistic processing. Evidence for the holistic processing hypothesis stems from studies which showed that facial parts such as eyes, noses, or mouths are better recognized from the full face than from a presentation of these parts in isolation (Tanaka & Farah, 1993).
In recent years, the holistic approach was often criticized for its confounded assumptions. Thus, the usage of the term *holistic* is often not independent from the configural processing hypothesis approach (Leder & Bruce, 2000b). Some authors assumed that configurations are the *basis* of the holistic hypothesis (Tanaka & Sengco, 1997), other scientists conceptualized holistic processing as an interactive influence of facial features on the perceptual representation of faces (Macho & Leder, 1998). Additionally, it seems that the so-called 'face holism' is not limited only to *faces* (see Donnelly & Davidoff, 1999), but can be expanded to many other complex objects and objects of expertise (Gauthier & Logothetis, 2000).
Recent studies have particularly investigated the role of configuration as a special processing in faces or as the binding element, which is the cause of the holistic whole-to-part superiority (Leder & Bruce, 2000b). Until today it is not at all clear what *holistic* exactly means. Some authors draw a pluralistic picture of 'holistic' mechanisms (e.g. Carey & Diamond, 1994), whereas others treat the many different 'holistic effects' as measures of the same underlying phenomenon (see Farah et al., 1998; Gauthier & Tarr, 2002). We can neither validate nor falsificate any 'holistic theory', until we have an exact definition of what holistic really means. This is probably the reason why so many results concerning the holistic processing hypothesis are rather arbitrary (Leder, 1999; Donnelly & Davidoff, 1999).

[16] For neural network accounts see Cottrell et al. (2001).
[17] Originally published in 1924.

3.1.3 Configural explanations

The following sections will illustrate the special role of configuration for the recognition of faces.

Special sensitivity to configural changes

Haig (1984) showed that we are sensitive to rather subtle configural facial differences. Relational changes between facial features in unfamiliar faces are sometimes detected at the visual acuity threshold level (Haig, 1984). Hosie, Ellis, and Haig (1988) found similar results using familiar faces. Kemp, McManus and Pigott (1990) used two-tone images and demonstrated that the high sensitivity for configural information is reduced in negative or inverted images. While these studies were primarily concerned with the perceptual level, Bruce, Doyle, Dench und Burton (1991) revealed a specialisation also for processing configural information at the level of memory processes (see also Carbon & Leder, subm.). In the Bruce et al. (1991) study, participants at test had to decide whether faces and houses were identical to the ones presented in a previous block or whether they were configurally altered. Although the alterations were smaller for faces than for houses, participants were more sensitive in detecting these alterations in faces. The authors therefore concluded that we are particularly specialized in configural facial changes (Young et al., 1987), but this sensitivity has to be developed over a long time period (Mondloch et al., 2002), probably until a human is matured. However, it is also true that infants have poor visual acuity and their cortex is exposed only to information of low spatial frequency which, for faces, specifies the global contour and location of features but little of the detail (Le Grand et al., 2001). Therefore, a rudimentary sort of configural processing must already be available for babies to be capable of differentiating between faces.

Importance of configural information

Real faces differ from each other in regard to their local components such as eyes, nose, and mouth (Rhodes, Brake, & Atkinson, 1993; Haig, 1986). Moreover, they differ with respect to texture and color (Hill & Bruce, 1996; Lee & Perrett, 1997) and in regard to higher-order—or configural—information (Haig, 1986; Diamond & Carey, 1986; Leder, Candrian, Huber, & Bruce, 2001; Leder & Bruce, 2000b). As we have seen above, even very small spatial displacements of face features (which approached the limit of visual acuity) could have a substantial effect on face matching (Haig, 1986). Thus, Galis and Mens (1986) argued that it is more likely that face recognition is based on higher-order invariances than on simple features. This idea is not new. More than a century ago, Francis Galton (1879) already suggested that the relationship among features of a face may be a more critical factor to its recognition than the contribution of any particular facial feature. Moreover, he suggested that configural information might be more vital to face recognition than the identification of individual features.

However, configural changes do not only come in play when all features are repositioned inside the facial context. As pointed out by some researchers, the changing of a single feature in a face also modifies the interrelation between the components and, therefore, the particular configuration of the face (Pomerantz & Pristach, 1989; Sergent, 1984b), or the 'microstructure' of components (Leder, 2001; Rhodes, 1988; Diamond & Carey, 1986; Leder & Bruce, 2000b; McKone et al., 2001). Additionally, every alteration of the form and size of one single feature changes the overall structure of spatial relations, because the distances between the edges of the features will be altered thereby (Leder et al., 2001; Leder & Carbon, subm.-a). Consequently, both sorts, local features and 'geometrical configural features' (Macho & Leder, 1998), are inherently involved in the perception of faces (Searcy & Bartlett, 1996). Therefore, it is practically impossible to distinguish them in natural faces.

Furthermore, it is assumed that expertise in face recognition, like the own-race advantage, is based especially on the encoding of configural information (Ellis et al., 1989; Rhodes, Tan,

Brake, & Taylor, 1989). Babies as being no face experts but rudimentary 'face recognizers' might perceive faces analytically, concentrating on particular parts of the face, such as eyes, nose, mouth or shape of the head, only putting the parts together at the processing stage (for the debate about analytical versus holistical processing, see section 3.2).

Locally configural features

A distinction can be made between two general types of configural processing, both of which assume the explicit encoding of face parts and their relations. On the one hand, *local configural* processing involves encoding of face parts critical for identification (eyes, nose, mouth, etc.) and their spatial relations independently from the face context. One the other hand, *global configural* processing involves encoding of critical parts and their spatial relations within face context. Thus, although both views stress the encoding of parts and their spatial relations, they differ in their assumptions about the role of the face context in configural processing.

In the preceding sections, the main focus was on global configurations, which are the spatial relations between the single features within the global facial context in terms of overall configural dependence. Another view was introduced by Leder and Bruce (2000b), who assumed that configural information is remembered locally and therefore independently of the surrounding context. Inversion deficits in the processing of relational information are not dependent on the context of the surrounding features, but are dependent on properties of constituent elements (Leder et al., 2001).

Inversion

There is clear evidence that upside-down faces are more difficult to recognize than upright faces (see for a review Valentine, 1988). This effect is well known as *inversion effect* and its investigation has a long tradition (e.g., Köhler, 1940; Hochberg & Galper, 1967; Goldstein, 1965). But Yin (1969) was the first researcher who showed that the effect of inversion is much stronger for *faces* than for other objects. Therefore, face recognition seems to be a 'special' process. This was replicated by many other studies, even with line drawings (Leder & Bruce, 2000a). It should be noted, though, that Valentine and Bruce (1988) failed to find a differential effect for mental rotation tasks.

Rock (1973; 1974) suggested that upside-down faces are difficult to recognize because the inversion disrupts the configural properties of the physiognomy (cf. Farah, Tanaka et al., 1995). Rock (1974) referred to Ernst Mach, who pointed out that the appearance of a square is quite different when it is rotated by 45°. After rotating a 'square' by 45°, it is rather recognized as being a 'rhombus' or 'diamond'. This hints at a very adaptive cognitive system that always strives to find meanings in objects. Maybe upside-down faces overtax a mechanism for correcting disoriented stimuli in the form of an 'encoding bottleneck' (Freire, Lee, & Symons, 2000). This might induce the cognitive apparatus to concentrate on isolated features, since all the components cannot be reversed simultaneously (Phillips & Rawles, 1979). Stürzel and Spillmann (2000) found results consistent with the idea that during the process of rotating a face from upright to upside down, the processing of the stimulus switches from a holistic to a more featural mode. The perception in the featural mode depends on individual features, with less regard to the face as a whole.

Comparisons between different classes of facial features revealed indeed that it is configuration which is disrupted by inversion (Bruce et al., 1991; Bartlett & Searcy, 1993; Leder & Bruce, 1998; Leder & Bruce, 2000b). These results suggest that relational information plays a more substantial role in the processing of upright faces than in that of inverted faces (Leder et al., 2001; Rhodes, 1993). Inversion effects are known to be relatively small for components presented in isolation, which indicates that even components include some but not much configural information (Leder et al., 2001), although results with swapped components sometimes revealed ambiguous results (Rhodes et al., 1993).

Nevertheless, inversion has many other important impacts on different aspects of face processing. For example, upright faces were more easily detectable than inverted faces (Purcell & Stewart, 1988), and attractiveness score ratings between faces decreased from upright to inverted orientation (Bäuml, 1994), suggesting that the faces became less discriminable by inversion. Moreover, socially relevant cues such as expression (Yin, 1970)[18] and grotesqueness (Krebs, 1998) of a face were much more difficult to evaluate for inverted faces. In addition, even audiovisual speech recognition is impaired when faces stand upside down. In a *McGurk paradigm* (McGurk & MacDonald, 1976), facial orientation did affect the accuracy of auditory speech report with incongruent audiovisual speech stimuli (Jordan & Bevan, 1997).
Inversion is also detectable in neuropsychological data. The N170 was found to be delayed and enhanced to inverted *faces* but not to other inverted *objects* (Rossion, Gauthier et al., 2002). Moreover, many object-sensitive cortical areas were found that are sensitive to inversion (Aguirre, Singh, & D'Esposito, 1999). In a monkey's brain, for example, more than 75% of the face-responding cells in the *superior temporal sulcus* and the *inferior temporal gyrus* were found to be orientation-sensitive (Perrett, Oram, & Ashbridge, 1998).
An explanation for the face inversion effect was proposed by Goldstein and Chance (1980). They argued that the disproportionate effect of inversion on face recognition was due to the development of a rigid face schema. Goldstein and Chance demonstrated that with repeated exposure to faces an increasingly rigid schema developed which enhanced the processing of faces of the type most normally encountered, but disadvantaged the processing of all faces outside the scope of the schema. Flin (1985) offered a very similar explanation for the inversion effect, suggesting that with growing experience with the upright position, an inverted image of any mono-oriented object would be perceived as increasingly odd. Diamond and Carey (1986) combined the two models into what they called *Goldstein-Flin hypothesis*. According to this hypothesis, face recognition may generally be more dependent on the encoding of spatial relationships between salient stimulus features than object recognition (Carey & Diamond, 1977; Diamond & Carey, 1986; Rhodes et al., 1993). The encoding of spatial-relational information in faces, on the other hand, may be disproportionately impaired by inversion (Searcy & Bartlett, 1996; Murray, Yong, & Rhodes, 2000). This theory has been partly validated by recent experimental data. Participants received successive daily practice sessions on the task of recognizing inverted faces (Haggbloom & Warnick, subm.). The results showed that sufficiently motivated participants became quite proficient at recognizing inverted faces, which had already been suggested by Bradshaw and Wallace (1971) and by Ellis (1975). Similar results with a rather different paradigm were obtained by Takane and Sergent (1983), who supposed that after 500 trials, subjects made use of an increased 'interactive' componential processing of inverted faces. Furthermore, it is assumed that upright faces are usually recognized with such an interactive processing mode.

Photographic negative

Photographic negative pictures of faces are quite comparable to inverted pictures, because although they both hold all information of a normal upright face, they are difficult to be processed. Bruce and Langton (1994) illustrated that photographic negatives are much harder to recognize than untransformed faces (see also Johnston, Hill, & Carman, 1992; Bruce & Langton, 1994; Kemp, Pike, White, & Musselman, 1996; Liu & Chaudhuri, 1998; Galper & Hochberg, 1971; Galper, 1970). However, what kind of information is lost when a photographical positive is transformed into a photographical negative? Is it a featural information loss, because now eyes will have bright pupils surrounded by dark regions, or is it a configural one—which was assumed by Lewis and Johnston (1997)—because our perception apparatus interprets other edges and corners than in the original variant. It could also be argued that the

[18] Contrary to this, Valentine and Bruce (1988) found no evidence for an interaction of orientation and expression detection.

configural information is not available *fluently* enough, which would also disturb the processing of facial expressions (Galper, 1970). Phillips (1972) discussed an alternative explanation. Maybe we mainly fail to recognize the gray levels of a complex object like a face, but are able to process pure black or white ones. We may have a special encoding of complex gray patterns as a whole, but not for the inverted pendants. However, Philips (1972) could find no support for this hypothesis by using *lith* photographs, in which only black and white color information was available. Therefore, an explanation based on the loss of configural information seems still to be very plausible and would parallel the findings for inverted faces.

Thatcher faces

Thompson (1980) first demonstrated a special perceptual effect by turning the eyes and mouth[19] region upside down, using a picture of the former British prime minister Margaret Thatcher. For this reason this kind of manipulation is nowadays called "thatcherisation" (Lewis & Johnston, 1997, p.225). The phenomenon associated with the *Thatcher illusion* is called facial anisotropy[20] and is very evident (in upright faces). If thatcherised faces are inverted, it is very hard to register any difference between the original and the Thatcher face (for an illustration see Figure 3-6). However, if the same pair of faces is presented upright—to test this effect, please turn this page upside down—the Thatcher version has a grotesque or even evil impression on the observer.

Figure 3-6: The Thatcher illusion demonstrated with a picture of the author. On the left side, the original, on the right side, the thatcherised version, where the mouth-and-eyes region is turned upside down (from Schwaninger et al., in press).

It seems probable that a key to the understanding of Thompson's (1980) effect might lie in the possibility that the mouth and eye features are to some extent processed separately, as distinct entities, embedded within the entire facial *Gestalt* (Parks, 1983). Bartlett and Searcy (1993) suggest that thatcherising a face changes configural coding without altering featural information. In the inverted form, these changes are not obvious since the faces are processed without configural information and the featural aspects are intact. This argument is similar to the one that has already been used by Köhler (1940) for inverted faces. He noted that facial expression almost disappears when a face is viewed upside down. The loss or decrease of configural information can also be demonstrated by a matching task. Bartlett and Searcy (1993) found

[19] Lewis and Johnston (1997) reported that inversion of the eyes alone is sufficient to produce the Thatcher illusion; teeth do not seem to be required, as the illusion persists in faces with their mouths shut.

[20] *Anisotropic*. Exhibiting properties with different values when measured along axes in different directions (Encyclopædia Britannica, 1999).

that face pairs of thatcherised and veridical faces were found much more similar when presented inverted than upright.
Rock (1974) reported data on the recognition of famous faces in photographs, presented upright, but with the observer's head at 0°, 45°, 90°, and 180° angles, relative to the vertical position. The data showed a monotonic decline in recognition performance as the observer's head rotated away from the vertical. Consistent data were found in the RT curves of Valentine and Bruce (1988). Therefore, they assumed that the process of *mental rotation* (Shepard & Metzler, 1971) was essentially the same for faces as it was for non-face stimuli. Folk and Luce (1987) found that especially mental rotation rates were slower for more complex stimuli than for less complex stimuli, but showed very similar patterns as such. Configural information like in a face is certainly more complex than isolated features, because they describe relationships between two or more objects, whereas features refer to single items (Rhodes et al., 1989). This was investigated in an indirect way by Phillips (1979), by demonstrating that recognition of the internal features of a face that consist mainly of configural information, was more disrupted by inversion than recognition of the external features.
Moreover, according to Leder et al. (2001), the disruption of the specialized processing of local features, such as shape of the eyes, contributes to the Thatcher effect. Therefore, the authors concluded that the effect is *not* caused by a disruption of processing relational information.
Another explanation comes from Parks (1983), who pointed out theoretically that the 'frame' of an upright face includes the relative position of the eyes and the mouth as well as the external features, and is therefore a very valid reference for detecting any alterations to the face. This sight was later contradicted by Valentine and Bruce (1985) with an experimental setting using inverted *inner* features in combination with an upright *outline*. Here, again a kind of Thatcher effect appeared, but in this case only with an *inverted* referential frame of a face outline.
Interestingly, Stürzel and Spillmann (2000) found by systematically varying the rotation angle of Thatcher faces that children below the age of three showed no surprise during the rotation of a thatcherised image, whereas children above this age exhibited surprise like adults did. This suggests that the specialized mode of recognition develops during early years of childhood. Children are, as discussed above, less sensitive to orientation of faces than are adults (Carey & Diamond, 1994). This reflects an increasing reliance on configural aspects of faces with increasing age and expertise. Young, Hellawell, and Hay (1987) showed that for adults the encoding of relations among facial parts is, indeed, sensitive to orientation.

3.2 Face Processing

In the preceding sections, models explaining the encoding and representing of models were discussed. Nevertheless, so far, little has been said about the *processing* underlying these models. The process of face recognition is a very complex one, and therefore we must focus on special details of it. Computational problems in vision are best understood at the early stage of visual processing, when the inputs and the goal of the computation are known (Ballard, Hinton, & Sejnowski, 1983). Thus, the next section will start with the first rudiments of face recognition, going on within the early vision and early processing (the first moments of processing).

Importance of early filtering and microgenesis

In what follows, the term *early vision/processing* is used roughly in the terminology introduced by Marr (1982), assuming that early vision is impervious to cognitive influences. Zenon Pylyshyn called such a processing *cognitively impenetrable* (Pylyshyn, 1999). Moreover, it defines the first moments of visual processing, where no strategic or volitional motivations can be involved (VanRullen & Thorpe, 2001). This is advantageous for studying face recogni-

tion processes, because with an impenetrability, it is much easier to localise and identify cognitive processes than those that are confounded with other cognition, or are too complex to be analysed. Oram and Földiák (1996) demonstrated that the stimulus complexity increases dramatically at successively higher levels.

There is a lot of information—most of it irrelevant to specific tasks—available even at the retinal level (Lennie, Trevarthen, Van Essen, & Wässle, 1990). However, this huge amount of complex visual information has to be broken down to be manageable by the cognitive apparatus. Therefore, it would make sense to minimize the information load at an early stage. One possibility to achieve this is to send only information about the restricted domains of the image, those, that are important to the organism. By introducing selectivity early, and thereby creating distinct classes of neurons that sample the stimulus domain selectively, the demands on central mechanisms are reduced. This can be done by identifying complex patterns as one region of interest, which can later on be analyzed with a higher priority and focus. This is the starting point of *microgenesis*, which will be discussed later. Through microgenesis, a complex pattern will be analyzed step by step, until it is finally recognized as a whole and stable entity.

The need for different processing modules

It has been suggested that face processing is mediated by a specialised module (Fodor, 1983; Yin, 1969; Farah, Wilson, Drain, & Tanaka, 1995; Nachson, 1995). According to Jerry Fodor, a module can be recognized as a mandatory, domain-specific, hardwired input system which performs innately determined operations, and therefore stands in contrast to the idea of a General Problem Solver (GPS) *in sensu* Newell and Simon (1972). Perceptual analysers are encapsulated from central knowledge and expectancies (Fodor, 1983; Treisman et al., 1990), that is, central cognitive processes are not shared and are cognitively impenetrable (Pylyshyn, 1999). Recently, the *encapsulated hypothesis* was criticized (Rhodes & Tremewan, 1993; Faulkner, Rhodes, Palermo, Pellicano, & Ferguson, 2002; Schyns & Oliva, 1999). For instance, Faulkner et al. (2002) tested the modularity hypothesis by examining whether discriminations between normal and distorted versions of famous faces can be primed, either by the name of an associated person (‘semantic context’) or by a valid cue as to the identity of the target face (‘expectancy’). The experimental data failed to support Fodor’s conjecture that face processing is encapsulated.

Marr and colleagues (1982) have proposed that perceptual information is carried by intensity changes which can occur at different scales and are best detected by analyzing each channel response separately. The premise of Marr’s approach is that intensity changes or edges carry enough perceptual information for a first visual report to be prepared for further analyzing. Thus, the first step in Marr’s decoding process is to locate the edges in the visual scene. Many visual features will not be represented in all channels, however. For example, small-scale channels are insensitive to gradual changes in illumination from one spatial position to another, while larger-scale channels are insensitive to fine detail, such as narrow lines and fine texture. Thus, the overall percept may be a composite of features, which are seen and processed separately at different scales, as well as features that activate channels in common. According to this view, the major task of the visual system is to form a symbolic description of the outside world. In order to form this description, the visual system has to rely initially on the retinal image, the 2D-projection of external visual appearance upon the back of the eye, which is called *Primal Sketch*. The Primal Sketch contains a description of the important structures in the image. The aim of this early symbolic process is to identify features of the image that are likely to be of importance in later scene-based analysis: features such as edges, bars, blobs, texture boundaries, corners, and so on. An early relative dominance of the components was observed for the stimuli whose components were closed figures (Kimchi, 2000). The Primal Sketch can be thought of as the internal representation of a cartoon of the image

(Shapley, Caelli, Grossberg, Morgan, & Rentschler, 1990), holding biologically important information.
One important stimulus class is that of faces. Cells found in the macaque *inferotemporal cortex* (IT), the visual area in the *ventral visual stream*—thought to have a key role in object recognition—are tuned to views of complex objects such as faces. These specialized neurons, called *higher-order hypercomplex cells*, discharge strongly to a face but very little or not at all to other objects in a reliable way, even if they are scaled or their position is changed (Riesenhuber & Poggio, 1999b). Thus, these *face detectors* operate similar to the 'hand detector' neurons in the macaque brain described in Gross, Rocha-Miranda, and Bender (1972). A hallmark of these cells is the robustness of their responses to stimulus transformations such as scale and position changes. These findings have now been complemented by studies in humans (Adolphs, 2001).

Microgenesis

The term *microgenesis*[21], first coined by Heinz Werner (1956), is an approximate translation of the German word *Aktualgenese,* respectively *Mikrogenese*, a creation by Friedrich Sander (1932). Microgenesis will refer to the sequence of events which are assumed to occur in the temporal period between the presentation of a stimulus and the formation of a single, relatively stabilized cognitive response (percept or thought) to this stimulus (Flavell & Draguns, 1957; Riffert, 1999). More specifically, the term will refer primarily to the pre-stages of extremely brief cognitive acts, e.g. the processes involved in immediately perceiving a simple visual or auditory stimulus.
As noted by Ellis (1983), several stages, or sub-processing units, compose the entire duration from input reception to response production (see also Sergent, 1986a; Thorpe & Fabre-Thorpe, 2001). These assumptions parallel Whitehead's (1985)[22] perceptional theory of a primary perception modus, in which he assumed a primary perception on which further forms are based on through microgenetic processes (see also Riffert, 1999).
In recent years, the microgenetic approach to pattern recognition has become quite popular (Sergent, 1986b; Watt, 1988). The main idea behind this approach is that different levels or aspects of the image become perceptually available at different moments of real time, while the accumulative process of percept development is going on (Bachmann, 1991). Although the percept appears to just be there—according to the microgenetic account—our visual representation of a scene is not arrived at in one step, but is rather built up incrementally. This assumption has strong theoretical *and* empirical support. Theoretical analyses (e.g., Ullman, 1984) have provided good reasons that some relational properties that hold between visual elements, such as the property of being inside or on the same contour, must be acquired serially by scanning a display. We also know from empirical studies that percepts are generally created by scanning and shifting the attentional focus and not by analyzing a stimulus in one slip. Even when attention may not be scanned, there is evidence that the achievement of simple percepts occurs in stages over a period of time (Calis, Sterenborg, & Maarse, 1984; Reynolds, 1981; Sekuler & Palmer, 1992). The 'visual brain' consists of several parallel multistage processing systems, each specialised in a given attribute such as color or motion (Bartels & Zeki, 1998). Each stage of a given system processes information at a distinct level of complexity.
Interestingly, in addition, processes were found that analyze stimuli in subsequent stages like in the *microgenesis* account assumed. For example, tuning to spatial frequency appears broader in brain areas *V2* and *V4* than in *V1* (Treisman et al., 1990). Furthermore, neurons in the inferotemporal cortex signal information about different aspects of a stimulus at different

[21] In the present work, the term *microgenesis* will not be used in the sense of cognitive development as used in Siegler and Crowley (1991).
[22] Originally published in 1927.

times, so that social information about a face is encoded at a later point in time than coarser information that simply distinguish a face from a non-face stimulus (Adolphs, 2001). Similar results have been found by Sugase, Yamane, Ueno, and Kawano (1999): coarse information sufficient for discriminating a face from a non-face (cf. Thorpe et al., 2001) was encoded relatively early, whereas more fine-grained information, regarding the emotion shown in the face, was encoded later. The finding is consistent with the idea that visual representations can include socially relevant information, perhaps via top-down inputs from structures such as the amygdala and the prefrontal cortex.
For example, visual categorization of a natural scene involves different mechanisms with different time courses: a perceptual, task-independent mechanism, followed by a task-related, category-independent process. Although average ERP responses reflect the visual category of the stimulus, shortly after visual processing has begun (75-80 ms), this difference is not correlated with the subject's behavior until 150 ms *post stimulum* (VanRullen & Thorpe, 2001).
If that is the case, then the following problem immediately arises. If the representation is built up incrementally, we need a mechanism for determining the correspondence between representations of individual elements across different stages of construction of the representation or across different periods of time. As we elaborate the representation by uncovering new properties of a dynamic scene, we need to know which individual objects in the current representation should be associated with the new information (Calis et al., 1984). This is the main point of the so-called *binding problem* (Vessel & Biederman, 2001), which will be discussed in detail later. Binding is thought of not as a process that generates a conscious experience, but rather one that brings different conscious experience together (Calis et al., 1984).

Functional model of face processing

The current most popular model of face processing is that suggested by Bruce and Young (1986)[23]. Bruce and Young (1986) propose three stages to facial processing: the *structural encoding* stage, the *face recognition* stage, and the *familiarity-decision* stage. In the first stage, the raw visual data enter the striate cortex and are processed as a face. Basic features of the image, such as orientation, color, and position, are processed to create the image of a face. In the second stage, the question of whether the face has a physiognomic invariant is asked. This means that the cognitive apparatus tests whether a face is familiar, what is generic and what is individual to this face. In the third stage, the familiar face is processed with all the episodic, semantic, and emotional memories that are necessary to identify the person. Only a few other competitive models are available (e.g., Hollis & Valentine, 2001). For an overview, see Bruce and Humphreys (1994). One of these alternatives, which investigates the microstructure of the Bruce and Young (1986) functional model of face recognition, is the *interactive activation and competition* network model (IAC) (Burton, Bruce, & Johnston, 1990), which discusses only the identification routes. It has a connectionist architecture and is implemented as a network model with interactive activation and competition. However, the type of connectionism mentioned here does not belong to the *PDP*[24] *class* (McClelland, Rumelhart, & The PDP Research Group, 1986), as the representations in the model are not distributed and do not incorporate a learning mechanism (with the exception of the IACL-model, which was modified later by Burton, 1994). Unfortunately, none of these models focuses on the *early* processing stages of faces.

In the next sections, different models of cognitive processing will be considered. The simplest one, the serial account, according to which every informational step is dependent on the result

[23] A review of this model will be found in Bredart and Bruyer (1994); for a critical review referring to clinical face recognition symptoms see Breen, Caine, and Coltheart (2000).
[24] PDP: *Parallel Distributed Processing*. For an overview of PDP logic and the connection with face recognition see Brunas, Young, and Ellis (1990). For a critical analysis see Fodor and Pylyshyn (1988).

of the preceding one, will be introduced first. Then the parallel model and some chimerical assumptions will be discussed. The chimerical accounts subsume serial as well as parallel processing logic under one approach.

3.2.1 Step by step: the serial approach

A recurring discussion in the study of the perception of human faces is the question, whether faces are processed in a *Gestalt* manner, in which the importance of configural relationships between features is stressed, or by means of a piecemeal analysis, that goes feature by feature in a serial way.

Sergent (1984b) reported strong evidences for a serial model of face processing. Analyses of reaction times indicated in all but one of her experiments that serial processing best described the operation inherent in the perception of faces (Sergent, 1984b). In accord with this, Bradshaw and Wallace (1971) required subjects to decide whether *Identikit* faces, presented in pairs, were the same or different ones. They found that the duration of such comparisons was best accounted for by a serial self-terminated mode of processing, in which each feature of a pair was compared until a difference was detected. Smith and Nielsen (1970) also reported results consistent with analytic/serial processing: The time needed to discriminate pairs of different schematic faces decreased as the number of differences between faces increased. Besides this, the serial process seems to proceed from the top of the face to the bottom (Smith & Nielsen, 1970; Sergent, 1982). The top part of the face (hairstyle and eye/eyebrow) is generally found to provide more information for later recognition than the bottom part. Shepherd et al. (1981, p.105) asked their subjects "What facial features draw your glance and hold attention?". The results were: eyes (62%), hair (22%), mouth (8%). Sergent (1984b) argued that this unequal contribution of the features to the recognition of a face could concur with the view that in facial perception an analytic mode of processing prevails.

Tversky and Krantz (1969), conducting a multidimensional scaling (MDS) analysis of dissimilarity judgment between pairs of schematic faces, suggested that the contribution of each facial component was independent of the values of any of the other components (see also Guttman, 1968; Kruskal, 1964; Kruskal & Wish, 1978). This is strong evidence that facial features do not interact with each other and that the formation of an overall impression of a face can be understood as being the sum of the independent sub-impressions obtained from single features (Sergent, 1984b). This view would be in accord with the assumption of serial processing. Macho and Leder (1998) did not find interactive feature processing either, but these findings might alternatively be explained by holistic processing (see section 3.1.2).

Tversky and Krantz (1969) suggested that subjects perform a serial self-terminating facial search, with the eyes and the face shape processed prioritised. They obtained their results by simple line-drawing faces. Smith and Nielsen (1970) used slightly more convincing schematic representations, which were more face-like. They concluded that dual processing mechanisms underlay recognition memory for faces, with 'same' judgments being determined by a holistic template comparison. Whereas, 'different' judgments resulting from a self-terminating serial process in which subjects appear to scan facial features sequentially from top to bottom. Hole (1994) explained why a 'same' judgment normally takes longer than a 'different' judgment, namely because subjects were engaging in a serial, feature-by-feature search for differences that terminated, once a difference was encountered. This is in accord with work by Caelli (1988) and Rentschler, Hübner and Caelli (1988), who suggested that the type of detectors, their tuning characteristics, and the use of detector responses by decision-making processes may be adaptive to the signal and task demands. Then, the "true *invariants* for visual information processing consist in the types of computations available" (Shapley et al., 1990, p.425), and this is task-dependent.

Additionally, it seems plausible, that faces are mentally scanned in a serial way, because the central bottleneck causes much of the slowing that occurs when two tasks are performed at the

same time (Rohrer & Pashler, in press; Ruthruff, Pashler, & Klaassen, 2001; Pashler, 2000). However, this evidence fits also in a parallel processing account.

Serial processing with a memory component

Accounts of visual search performance typically assume (implicitly or explicitly) that subjects search through the items in the display one by one, without retracing their steps, or, in terms of probability theory, that visual search proceeds by sampling without replacement. To be able to do this, there has to be some memory mechanism that keeps track of previously attended locations. This assumption of memory-driven search is a central principle of the standard self-terminating serial processing model (Sternberg, 1966), which has been assumed in almost all models of visual search performance with a serial processing component (e.g., Grossberg, Mingolla, & Ross, 1994; Wolfe, 1994; Treisman & Gelade, 1980; Schneider & Shiffrin, 1977; Treisman & Sato, 1990; Horowitz & Wolfe, 2001).

The empirical support for memory-driven search is surprisingly thin, given its widespread acceptance in models of search. The leading hypothesis, explaining this memory-driven form of search, was formulated by Posner and Cohen (1984). They proposed that *inhibition of return* (*IOR*)[25] served to prevent attention from being deployed to rejected distractors (Tipper, Driver, & Weaver, 1991)[26]. This view has been opposed, recently. Horowitz and Wolfe (2001) claimed that visual search is actually 'memory-free', by which they meant that no record was being kept of the deployments of attention during a search. Their experiments involved a 'dynamic search' condition, in which they disrupted this hypothetical memory for deployments of attention during a trial, by replotting all items at new, randomly chosen locations, every 100 ms. This is much closer to reality, where the contents of a scene may change rapidly or the initial identification of an object may be faulty (cf. Horowitz & Wolfe, 1998). Other researchers were able to demonstrate that *IOR* does not seem to rely on covert attentional orienting because *IOR* is just as strong at each of two separate simultaneously cued locations as at a single cued location (Lambert & Hockey, 1991).

Evidences against strict serial processing

To distinguish between serial and parallel processes has a long history (Neisser, 1967; Kinchla, 1974; Schneider & Shiffrin, 1977). The strict serial/parallel dichotomy is a useful, but potentially dangerous fiction. It is no longer part of the *Feature Integration Theory* of the Treisman group (see Treisman, 1993; Treisman & Sato, 1990), and it is nowadays explicitly rejected by various other models of visual search (Humphreys & Müller, 1993; Wolfe, 1994; Wolfe, 1996; Wolfe, Cave, & Franzel, 1989). Sergent (1984b) demonstrated that the time to detect a difference between faces differing on two dimensions was shorter than the time to detect differences on the fastest dimensions alone, which stands against a strict serial processing hypothesis. Other researchers were able to show that only scrambled faces were categorised following a (self-terminating) serial search of facial features (Donnelly, Humphreys, & Sawyer, 1994). This was not demonstrable for both upright and inverted representations. The recognition of such faces rather follows a matching process, going on in parallel with a stored mental representation of a face (Donnelly et al., 1994).

Further criticism against strict serial models is based on theoretical thoughts about the nature of processing. The pattern of results produced by a serial self-terminating search can also be produced by a variety of limited-capacity parallel models (Kinchla, 1974; Ratcliff, 1978; Ward & McClelland, 1989). The idea of a limited capacity parallel model is that all items in the display are processed, at once. Search terminates when one item crosses the 'yes'-threshold or when all items cross the 'no'-threshold (Wolfe, 1996). Therefore, it is not easy to

[25] *Inhibition of return*. The tendency not to return to recently attended locations in orienting experiments.

[26] Nowadays, the theory of *IOR* was corrected to be an *object*-based rather than a *space*-based effect (Müller & von Mühlenen, 2000).

definitely make a decision whether empirical processing data can really be identified as mirroring serial or parallel processes.

Attention capturing

The construct of salience is closely related to that of attention capturing. In the empirical part of the present work, the salience of several features in a face context will be manipulated. Therefore, it is important to understand the attention-capturing mechanisms and their impact on face recognition.

Schneider and Shiffrin defined selective attention as "control of information processing so that a sensory input is perceived or remembered better in one situation than another" (Schneider & Shiffrin, 1977, p.4). The concept of selective attention is predicated on the assumption that attention resources are limited. Therefore, to prevent a processing capacity overload (Tsotsos, 1995), special attention must be given to the subset of information which should be the most relevant one (Chun & Wolfe, 2001). Attention is mostly assumed to be processed sequentially and to focus on the salient image locations (Treisman & Gelade, 1980; Koch & Ullman, 1985).

Furthermore, Schneider and Shiffrin (1977) proposed two quantitatively and qualitatively distinct processes to account for their results (for a possible schematic model of such a system, see Heinke & Gross, 1994). They proposed, on the one hand, a controlled search, on the other hand, an automatic detection mechanism. Moreover, they assumed that controlled search is a serial process in which a matching decision occurs after comparison of each item in the display to memory set items. In contrast to that, the automatic detection operates in parallel and independent of attention. Automatic processes do not only require no attention, they even do not use up short-term memory capacity. In addition, automatic processes are assumed not to be very flexible—they are barely modifiable and are difficult to stop, once initiated. Schneider and Shiffrin (1977) do not seem to make a distinction between automatic detection of featural information and automatic detection that requires more training. This is noteworthy, because the research group around Treisman (e.g., Treisman & Gelade, 1980) have argued that while simple featural search and trained automatic responding appear similar, they have different origins of independence from attention. Automatic detection achieves its non-attentive status from earlier attention through practice. By contrast, automatic feature detection is an innate process supported by populations of feature detectors. Logan (1992) made a further distinction between these processes by arguing that automatic detection is a post-attentive process that is dependent on attention.

Why is attention so central to perception?

In some of their recent papers, O'Regan and Rensink (e.g., O'Regan, 1992; Rensink, O'Regan, & Clark, 1997) proposed that vision is very poor when we do not attend to the critical objects. To put it more radically, they postulate that attention is not only needed to see changes, but to see *anything at all* (O'Regan, 1999).

As mentioned by Pashler (1995), it is not clear whether it actually makes sense to postulate a model of vision in which observers see everything in a scene, but when a change occurs, they cannot see it unless they are attending to it. Such a view would assume that we represent an inner picture of a current scene 'that we see', but cannot access any information of this picture-like representation unless we focus our attention on a part of it. Philosophically argued, it is more coherent theoretically to change this model into a less sophisticated one with fewer assumptions, where nothing is seen unless it is attended to. This is also the view taken by Mack and Rock (1998) in the context of their studies of *inattentional blindness*.

Spotlight metaphor

It has often been assumed that there is a single focus of attention, representing the current locus of visual processing, which can be moved across the visual field independently of eye

movements (Pylyshyn et al., 1994). The precise details of this, the so-called *spotlight metaphor*, vary among investigators, although it is universally thought to apply to a single contiguous spatial region. Moreover, it is believed that even if attention is unitary and spatially focused, there must be an additional primitive mechanism for simultaneously indexing several places in a visual field to enable excluding already triggered points for further processing (e.g., Pylyshyn, 1989, 2001; Burkell & Pylyshyn, 1997; Sears & Pylyshyn, 2000). One possibility for such functioning is formulated by the *FINST* hypothesis (Pylyshyn, 1989; Trick & Pylyshyn, 1993), which operates on early maps, such as Marr's (1982) *Primal Sketch*, the *Feature Map* of Treisman and Gelade (1980), or Wolfe's (1989) *Activation Map*.
Also known as *attention window* (see Kosslyn, 1994), the spotlight metaphor can select a contiguous set of points for deep processing. This may help to identify objects at different parts of the images. Observers perceive this as rapid eye movements. Before an eye movement is made, attention is covertly shifted to the location of the object of interest (Theeuwes, Kramer, Hahn, & Irwin, 1998). The eye movement, required to bring the fovea into register with the peripheral target, is called saccade. Saccades usually take 200-300 ms (Julesz, 1981a; Kosslyn, 1994) before they are initiated, with a latency that is highly dependent on the stimulus conditions and the state of attention (Treisman et al., 1990). Moreover, saccades are ballistic motions of up to 900°/s during which visual information processing is severely limited (Volkmann, 1986). But such intentional saccades only come into play when there is overt attention (Posner, 1980; Deubel & Schneider, 1996). However, with given PTs of fewer than 200 ms, they are not possible. Therefore, to exclude saccadic movements, the PTs will be limited to 200 ms for the experiments of the present work.

Saliency Maps

Koch and Ullman (1985) have hypothesized a centralized two-dimensional *saliency map* that can provide an efficient control strategy for the strategic positioning of attention on the basis of bottom-up cues. The input image is decomposed through several pre-attentive feature detection mechanisms, which operate in parallel over the entire visual scene. Neurons in the feature maps encode the spatial contrast in each of those feature channels, which are highly specialized to special properties of a stimulus (see also Kahneman & Treisman, 1992), and compete among each other for the highest salience. The most salient features are then combined into a unique saliency map, which encodes for saliency irrespective of the feature channel in which the stimuli appeared salient. After that, the saliency map is sequentially scanned by attention in a winner-takes-it-all manner to identify the highest saliency at any given time and by an *inhibition of return* mechanism, which suppresses the last attended location from the saliency map, to prevent redundancy. Therefore, such models are also called *prioritizing models.* To attune this model to very common stimuli like faces, a top-down attentional bias and training can modulate most stages of this bottom-up model (see Heinke & Gross, 1994).
A rather similar idea is incorporated in Wolfe's *Guided Search model* (Wolfe et al., 1989; Wolfe, 1994; Wolfe & Gancarz, 1999; Wolfe, 2001), although this model incorporates an additional mechanism for top-down influences on selection based on preactivation of goal-relevant feature maps. Contrary to this, Watson and Humphreys (1997) proposed a template-based inhibition of feature maps (e.g. Kim & Vase, 1994) linked to old objects and providing a mechanism for making new objects salient. Hence, a model like that of Watson and Humphreys (1997) might better be called a *deprioritizing model* and not a classical, positively termed *prioritizing model* (as described in Yantis & Jonides, 1984).
Besides some electrophysiological support (e.g., Itti & Koch, 2001) for the saliency map model and the intuitive character of it, a number of visual processing tasks are clearly performed too fast for such a costly strategy to be employed (VanRullen, in press). Therefore, it was recently suggested that an implicit representation of saliency can be best encoded in the relative times of the first spikes fired in a given neuronal population without any awareness, propagated in a feed-forward fashion (Delorme & Thorpe, 2001). A simulator of modeling

large populations of synchronously spiking neurons is for instance *SpikeNET* (VanRullen, Gautrais, Delorme, & Thorpe, 1998; Delorme, Gautrais, VanRullen, & Thorpe, 1999; Thorpe, 2001).

Other criticism of saliency maps comes from researchers who are skeptical as to whether the commonly used experimental setting of a static visual search field is at least approximately realistic. In everyday life, we look more often at dynamic scenes than at static ones. For that reason it seems implausible that we should use such a cost-intensive and slow mechanism to trigger display slots that have been attended before. A random scan is a much quicker way to find items in a display than a systematic search, as deliberate movements of attention are significantly slower because of an internal limit on the speed of volitional commands. "Anarchy is faster than order in this case" (Wolfe, Alvaraz, & Horowitz, 2000, p.406). It is speculated that the order of search, even of eye movements, is influenced by stimulus salience and eccentricity, but is otherwise random through the set of salient loci. Indeed, an 'anarchic' search modus is very meaningful if the search constellation alters very quickly (Horowitz & Wolfe, 1998), whereas any memory-driven search mechanisms would be disastrous.

However, what is common to all these models dealing with saliency? Research of feature saliency in face recognition provides support for an analytic mode of processing. Moreover, the most salient features are not only the easiest to discriminate but are also the features that are processed first (Walker-Smith, 1978). Evidence that the various facial features are not equally salient would then confirm that they are not processed simultaneously (Sergent, 1984b). There is a large body of literature suggesting that facial features are not processed equally effectively in assisting face recognition (see Shepherd et al., 1981; Davies et al., 1977; for reviews see Ellis, 1975), and it seems that these differently salient features are also processed more or less quickly.

Face-specific attention

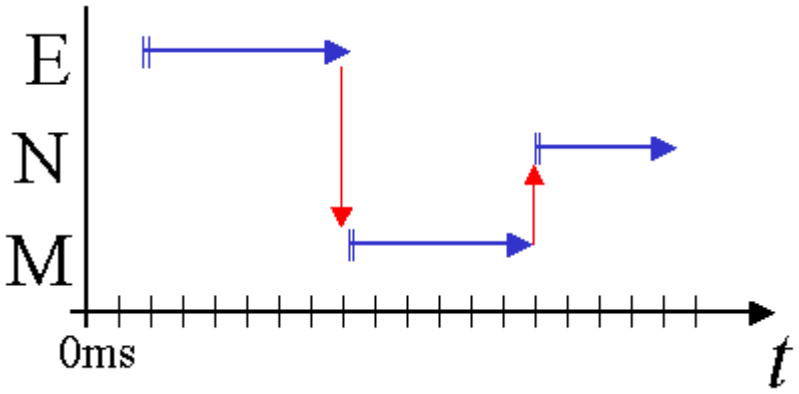
Faces seem to have a special capacity to attract attention for further processing. Attention-capturing properties should become most apparent in situations where real face stimuli are competing with other real objects for attention (Ro et al., 2001). This can be demonstrated by the phenomenon that attention will be spontaneously cued to a face even in the absence of any instruction (Levin & Simons, 1997).

The advantage of such a guiding mechanism seems obvious. Few other imaginable objects contain more socially and biologically relevant information than faces. This was validated by Rensink et al. (1997), who found that changes to important objects are detected much faster within a *change blindness* paradigm than changes to objects of marginal interest. Thus, in 50% of the cases where an observer looked at a marginal-interest location in the moment it changed, this alteration was not noticed, even though the eye was located less than one degree from the changing position (O'Regan, 1999).

Typical serial processing model for facial features

As a summary of this section, a typical serial processing model for facial feature analysis is illustrated in Figure 3-7.

Figure 3-7: Typical serial processing model. The processing sequence of the features is the following. First, the eyes (E) are processed, then the mouth (M), and finally the nose (N). The red arrows indicate that the starting of feature processing is dependent from the termination of the processing of the preceding feature.



3.2.2 Parallel processing

The computation collectively done in the brain is much faster than the fastest computer existing today. Because biological neural pathways are generally responding rather slowly, this high performance is attributed to the parallelism of the computation (Treisman, 1996; Singer & Gray, 1995). Work in computer science (e.g., Ballard et al., 1983) suggested that breaking up a task into components that can be computed separately is a practicable means of coping with slow computing elements. Marr (1982) called this the 'principle of modular design'. The important question concerning such a modularity is, how should perceptual tasks be segmented? Visual scientists have devised several techniques to show how perceptual tasks might be decomposed into component problems that are solved relatively independently.

Treisman and Gormican (1988) and Lennie et al. (1990) discussed a number of such techniques to decompose complex processes into their inherent modules; the same was done for computer vision (Marr, 1982). Lennie et al. (1990) proposed two *modi* of vision, which are assumed to be processed in parallel. On the one hand, vision has to play an important role in orienting the subject in space and guiding its larger movements. This kind of vision is dependent on the peripheral visual field and is particularly sensitive to motion. On the other hand, focal vision is used for the detailed examination and identification of objects and is associated with the fovea and exploratory eye movements. These two forms of vision, often characterized as 'where' and 'what' vision, are linked to two major projections of optic nerve fibers, to the midbrain and to the thalamus (Lennie et al., 1990). The 'where' vision corresponds thereby with the parallel processing account and the 'what' vision corresponds with the serial processing model.

Evidence for early parallel processing stems from visual search tasks, in which targets that 'pop out' against the background of distractors have to be discovered. Such targets are detected equally quickly regardless of the number of distractors that are present. Fast and spatially parallel detection is taken as evidence that the features in question are coded early in the visual process (Treisman et al., 1990). Studies by Treisman and colleagues showed that the human visual system is extremely good at identifying certain characteristics without search. They have assumed that features are projected into separate spaces that can each contain one feature plus positional dimensions (Treisman & Gormican, 1988; Treisman & Gelade, 1980; Treisman et al., 1990; Treisman, 1982).

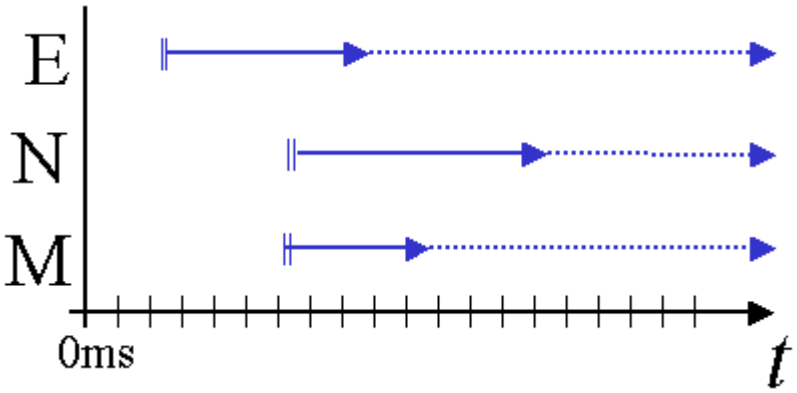
More recently, in Treisman and Gormican (1988) the assumed strict dichotomy of features being detected either in parallel or in serial (e.g., Treisman & Gelade, 1980) has been abolished. It seems more plausible that parallel and serial represent two ends of a spectrum. 'More' and 'less' are also encoded on this spectrum, not just 'present' and 'absent'. More-

over, the amount of differentiation between the target and the distractors for a given feature will affect search time (Treisman, Vieira, & Hayes, 1992).
Another evidence comes from reaction time (RT) data. Because facial features are not equally salient, and therefore are not processed at the same rate, a purely analytic mode of comparison would predict the following RT pattern. The fastest RT to discriminate between faces could not be faster than latencies to discriminate between faces differing in the more salient features. But such a pattern did not prevail with upright faces, indicating that information from several features was combined and interacted before a decision was made, resulting in a decrease in latency (Sergent, 1984b). Recently, Lamberts and colleagues have also presented strong evidence for the parallel independent sampling of an object's parts in the earliest stages of categorization (Lamberts & Freeman, 1999b, 1999a; Lamberts, 1998).
Additionally, it could be demonstrated that identity and expression recognition operate independently of one another within the paradigm of *composite effects* (Calder, Young, Keane, & Dean, 2000). Similar results were found by Young et al. (1986) by comparing facial identity matching data with facial expression matching data of familiar and unfamiliar faces. This is also in accord with the assumption of independent pathways of identity and expression processing through the Bruce and Young (1986) model. Nevertheless, in a recent work of Hänggi (in prep.), this independence could not be found. In this study, the cognitive effort increased for an expression decision task with a simultaneously ongoing identity-matching task.

Typical parallel processing model for facial features

As a summary of this section, a typical parallel processing model for facial feature analysis is illustrated in Figure 3-8.

Figure 3-8: Typical parallel processing model. The processing of the single features is not interdependent. In this example, the eyes (E) are processed first, but this processing does not have an influence on the succeeding processing of the other features (N: nose, M:mouth).



3.2.3 Chimerical processing: Serial and parallel

Matthews (1978) has proposed that the perception of faces involves a mixture of parallel and serial processing. Using *Identikit* faces and 'same-different' judgments of simultaneously presented faces, he found that changes to either hair, eyes, or chin were detected equally fast and more rapidly than changes to eyebrows, nose, and mouth, the latter with increasing latency. Matthews (1978) interpreted these results in terms of a dual processing strategy, but there are two checks that should have been made on the data before suggesting that the hair, eyes and chin were processed in parallel. It is indeed conceivable that equal RTs to these feature changes are an averaging artefact (Sergent, 1984b). For example Walker-Smith et al. (1977) have shown that subjects differ in strategies they use in scanning a face. Some participants started at the centre, others at the top, and still others at the bottom. Assuming different scanning strategies, equal RTs may then be consistent with serial processing since averaging latencies across subjects may conceal individual differences. It is regrettable that Matthews (1978) did not examine patterns of RT distribution across subjects. A similar check on individual patterns of results would have been necessary to ensure that each subject was using a parallel processing strategy, since each subject may have used different serial strategies over the trials, which would not be apparent after averaging data within each condition. Therefore, Sergent (1984b) agued that unless an examination of RT distribution between and within each condition for each single feature difference reveals a unimodal distribution, the use of averaged data to infer the nature of the underlying processes is likely to be equivocal and misleading. This suggests that Matthews' (1978) interpretation may not be validly grounded (Sergent, 1984b).

Psychophysical evidence for a two-stage model has shown the existence of parallel and serial components of the visual search process (Harner & Gaudiano, 1997; cf. Bradshaw & Wallace, 1971). First, in a *pre-attentive stage*, parallel processing extracts local features. In a second *attention stage*, serial processing performs a more detailed investigation in areas of specific interest. Such a model can explain how a target with unique features 'pops-out' as well as how the time to find a target that shares features with other items increases with the number of distractors. Additional evidence for the proposed pre-attentive stage is delivered by neuropsychological studies. The visual system extracts simple features, such as orientation and color, automatically and in parallel over the whole visual scene: columns of neurons in brain area *V1* are tuned to different orientations. Blobs in *V1* are tuned to different colors. Their topographic organization in *V1* forms numerous feature maps. There is equally evidence for the assumed second stage of serial attentive selection. Attention draws the eye to salient features or *areas of interest.* More detailed processing occurs in the high-resolution fovea. Because the fovea can process only a small area[27] at a time (see further details in Fulton, 2000), this type of processing must be serial (Harner & Gaudiano, 1997). Similar two-stage accounts were made by many other psychological researchers (e.g., Treisman & Gelade, 1980; Wolfe et al., 1989; Heinke & Gross, 1994) and in the field of automated recognition systems (Scassellati, 1998).

Popular two-stage models

The most popular two-stage model is presumably the *Feature Integration Theory* (FIT) by Treisman and colleagues (e.g., Treisman & Gelade, 1980). Once, objects have been delineated by early processes of feature-based segregation and boundary formation, they must be identified as known individuals or instances of familiar category. Identification is postulated to involve the use of two different strategies. The first is based on the use of particular, salient cues that directly label the object rather than on elaborate processing of all its parts. The second

[27] Some authors define this area as being smaller than 1º of visual angle (Fulton, 2000); others define it as being about 2º of visual angle (Minut et al., 2000).

strategy allows us to extract relationships between features (i.e., what is above, what is below). It is the second process that enables us to recognize the face in more detail and to differentiate faces by virtue of single features. There is evidence that the two processes can be dissociated in visual agnosia (Treisman et al., 1990) and in certain visual illusions such as the *Thatcher illusion* (Thompson, 1980).
The character of the two stages is the following. Single features are registered early, automatically, and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention (Treisman & Gelade, 1980) and therefore are processed serially. The first parallel stage functions only for low-level features, which are called *pop-up properties*. In contrast, the more complex and conjunctive features are being handled by a serial process.
A second, very frequently cited model for two-stage processing is the *Guided Search model*, which is built around the idea that the continuum can be explained by an early parallel mechanism working in tandem with a later serial mechanism (Wolfe et al., 1989; Wolfe, 1994; Wolfe & Gancarz, 1999; Wolfe, 2001; Kinchla, 1974)[28]. Thus, it breaks down the frequently used distinction between serial and parallel processes. The origin of the interplay between both processing classes will be found in Neisser (1967), who proposed a model with preattentive stages of vision that are characterized by parallel processing of basic features. The parallel processing is followed by a bottleneck after which processing is essentially serial. The selection of items for serial processing is under attentional control. Similar to the assumptions of Treisman and Gelade (1980), these two-stage models hold that the explicit knowledge of the relationship of features to each other requires serial processing. The variation in the efficiency of search is determined by the ability of preattentive, parallel processes to *guide* attention toward candidate targets or away from likely distractors. Therefore these models are often noted as *Guided Search* models (Chun & Wolfe, 2001).

Analytic vs. holistic

It is a well-explored fact that the characteristic right hemisphere mode is a holistic processing, whereas the left hemisphere seems to process more in an analytic way when recognizing faces (Paller, Parsons, Grabowecky, Mayer, & Rao, under review; Rhodes, 1993). It was also found that there might be other functional brain asymmetries[29]. For example, at birth the right hemisphere might be specialized in processing face-like patterns (Valenza et al., 1996). This was also shown for adults. It was found that the early visuospatial processing as well as the creation and comparison of facial representation appear to be carried out more efficiently by the right hemisphere (for a critical view on this point see Sergent, 1986a), whereas comparisons based on discrete, nameable features of faces may yield a left-hemisphere advantage (Rhodes, 1985).
Neuropsychological evidences were mirrored by many cognitive-psychological experiments. For instance, children seem to rely strongly on analytic face recognition processing. It is easy to hoax their face recognition performance by adding specific known features or attributes like a scarf, glasses, a moustache or a hat to a face (Diamond & Carey, 1977; Carey & Diamond, 1977; Bartrip, Morton, & De Schonen, 2001). Very young children (eight months old) even focused and relied on single features like the mouth, which is a particularly important feature for them (Schwarzer, Huber, Korell, & Zauner, 2002; Schwarzer & Massaro, 2001). Therefore, age seems to be a central variable in part-whole perception as well as in analytic-to-holistic processing (Goldstein & Mackenberg, 1966). Recently, additional support for this was given by ERP data (Chaby, Jemel, George, Renault, & Fiori, 2001).

[28] For a comparison with parallel processing models see Eckstein, Beutter, Bartroff, and Stone (1999).
[29] Investigators of face recognition impairments caused by brain injury have revealed a number of important features of the human face processing system (Young, 1992). One of the first of these to be shown was the involvement of right cerebral hemisphere, which was suggested in the 1950s.

Matthew's (1978) experimental data, based on reaction times, postulate a dual processing strategy for face recognition. The dual process combines an analytic analysis based on hierarchical feature salience with an overall holistic, that is, configural analysis of the face. Research on eye movements is consistent with this explanation. It has been reported that participants fixate first on the configural features (i.e. hair, eyes, and chin) before analyzing other features (Walker-Smith et al., 1977). Bruce (1988) also argued for the dual processing explanation. Bruce said that an initial holistic analysis (gathering overall information about the face) directs the succeeding featural processing of the details of the face. The two processes continue in parallel to yield more information about a face. Diamond and Carey (1986) have argued that face expertise enables parallel coding of configural *and* isolated features. This was tested by comparing the effect of inversion of a face on recognition of 'own race' (high expertise) and 'other race' (low expertise) faces (Zebrowitz, Montepare, & Lee, 1993). Use of configural information was associated with a larger inversion effect than use of isolated features, and therefore inversion produced lower recognition for own-race than for other-race faces (Rhodes et al., 1989; Zebrowitz et al., 1993). This was contrary to the results of Valentine and Bruce (1986c), but their design had the specific weakness of unequal presentation times (see for details Rhodes et al., 1989).

There is some parallelism in computer-science to psychological face recognition theories. In the research and development of automatic face recognition systems, there are also two basic classes of face localization and recognition methods: an attribute-based (or feature-based), and a template-based (or holistic-based) account. Using an attribute-based account, specific attributes like elliptic shape, certain colors, or facial features are searched for in the image. Such methods are usually quite fast, but less robust. Using a template-based technique, whole *Gestalten* are given as examples with the goal to find out objects similar to these examples in a *ganzheitlich* way (see Wertheimer, 1938). This method needs much more effort for development, but leads to better results.

To summarize these findings, holistic processing is a much more computationally demanding and sophisticated kind of recognition, but is therefore also a more robust one compared to attribute-based methods (Ahlberg, 1999).

Global vs. local precedence

There is much psychophysical evidence for spatial-frequency channels in the recognition of complex objects and faces (Graham, 1981). The notion of independent perceptual channels came into vogue again especially with the work of Julesz (1981a) and his postulation of spatial-frequency analyzers in vision (original work made by Neisser, 1967; Flavell & Draguns, 1957). Just some years before Julesz's findings, Navon (1977) wrote the seminal paper '*Forest before trees!*', which traced as main hypothesis the concept of *global precedence*. According to this hypothesis, the visual system processes whole pattern information before information about pattern substructures, whereby globality is equaled with low-resolution information. These results were paralleled by many other researchers (e.g., Hucka & Kaplan, 1996). Hughes, Layton, Baird and Lester (1984) found global precedence involved in the early visual system but could also improve the hypothesis by demonstrating that not globality as such, but that the low spatial information is the key issue of precedence. They found that patterns that normally yielded global precedence, when filtered to remove low spatial frequencies, failed to yield a global advantage and instead led to local precedence! Moreover, Paquet and Merikle (1984) showed that global interference is affected by exposure duration in the processing of a hierarchical structure. They showed that only global-to-local interference occurred at short exposure durations. In contrast, global-to-local as well as local-to-global interference was observed at long exposure durations, which could be an indicator that the effect of exposure duration on global interference is explained not only by spatial-frequency channels, but also by attentional shift (Hibi, Takeda, & Yagi, 2002). When presented early, large-scale fragments tended to facilitate identification more than small-scale fragments (Sanocki, 1993;

Sanocki, 2001). In these experiments, the pure size was an effective definition of scale. The results are consistent with the idea of a microgenesis that varies with the density of pattern elements (e.g. Kimchi, 1998, 2000). Other researchers also argue for a predominantly coarse-to-fine process in vision (Schyns & Oliva, 1994; Oliva & Schyns, 1997; Delorme, Richard et al., 1999), which is contrary to Biederman's (1987) *recognition by components* (RBC) theory. RBC theory assumes that local objects (*geons*) are recognized earlier than global structures. Schyns and Oliva (1994) used *hybrid* images, which are composed of a low spatial-frequency content of one scene and a high-spatial frequency content of another scene. The results implied that at short exposure durations, the low-frequency components, which represent the coarse spatial structure, had a greater impact on perceptual processing than the high-frequency content of the hybrid. But if different spatial scales transmit different information about the input, an identical scene might be flexibly encoded and perceived at the scale that optimizes information for the considered task (Oliva & Schyns, 1997).
Thus, recognition occurs at both coarse and fine spatial scales (Schyns & Oliva, 1994; Uttal, 2001), and no particular range of frequency is absolutely essential for face perception (Sergent, 1986b), but all the frequencies are not equally useful. Which frequency bands are important depends on the task. For example, low-dimensional representation is best for identifying the physical categories of a face, like sex, but it is sub-optimal for recognizing the faces, for instance discriminating a known from an unknown face (O'Toole, Abdi, Deffenbacher, & Valentin, 1993). Therefore, different bands of spatial frequency are used to carry out different operations that can thus take place in parallel and increase processing efficiency (Sergent, 1989, 1986a; De Valois & De Valois, 1980). Different bands can also be extracted by PCA methods (e.g. Valentin et al., 1998; Hancock, Burton, & Bruce, 1996).
However, different frequency scales do not only result in different and specific recognition performances. Moreover, humans can react to low-frequency stimuli more rapidly than to high-frequency stimuli (e.g., Parker & Dutch, 1987). Low-pass faces yielded longer latencies overall, and were processed significantly faster in *left visual field* (LVF) than in *right visual field* (RVF) presentations. These results confirm the greater capacity of the right hemisphere (RH), compared to the left hemisphere (LH), to operate on low-frequency contents of faces (Sergent, 1986a). Thus, both hemispheres are competent at processing faces, but each is a specialist (Sergent, 1986a), here for specific spatial frequency bands.
However, why should global or exterior qualities be processed earlier than local ones? Waltz (1975) demonstrated that exterior edges and forms are more reliable starting points for recognition mechanisms. Contours and global structures are better to identify the *figure-ground* relationship (Baylis & Cale, 2001; Julesz, 1981a). Additionally, when the space extensions of a figure are realized, cognitive resources and attentional effort for later processes can be better estimated and organized. Besides this, global structures are larger, therefore better recognized even from a greater distance, and are more efficient for recognizing the object *identity*.
In addition, because we know that attention cannot be simultaneously committed to multiple non-contiguous spatial channels (Posner, Snyder, & Davidson, 1980; Hoffman & Nelson, 1981), it seems clear that attention cannot be summoned to all abrupt visual events, at once (Yantis & Jonides, 1984), and therefore is plausible to go on in stages or microgenetic steps.
Neuropsychological accounts can explain the global-to-local transfer by the *spreading activation theory* (Anderson, 1983) and the fact that the complex visual information, given by the presentation of a face, is not only processed by different processors but that these processors may also take different time to complete operations (Breitmeyer & Ganz, 1976). The sum of these facts must evidently cause a sequence of different identification stages, but it is not known whether these steps are necessarily dependent on each other and happen as a series.

The binding problem

Von der Malsburg (1981) formulated the *binding problem*. His claim was that visual representations based on spatially invariant feature detectors were ambiguous, because if there are

general mechanisms to detect and identify single features, it is still unknown and vague how these features should be integrated to a whole *Gestalt*. Neuronal systems have to solve immensely complex combinatorial problems, and they require efficient binding mechanisms in order to generate representations of perceptual objects and movements (Tsotsos, 1990). As Patricia Churchland has observed, the question as to how the nervous system integrates information is essential (Churchland & Sejnowski, 1992; Churchland, 1986)[30]. A spatial binding problem arises, for example, when each detector in the visual representation reports only the *presence* of some elemental object attribute but not its spatial location (Mel & Fiser, 2000). The binding problem is best characterized by the question of how distributed sets of features can represent multiple objects without confusing the features of the individual objects (Stoet & Hommel, 2002; Kitcher, 1990). Therefore, object representation necessarily has to be processed through at least two phases: feature *activation* and feature *integration* (Stoet & Hommel, 2002). One possible theoretical realization of this requirement is a hierarchical feedforward model (Riesenhuber & Poggio, 1999a) from simple cells all the way up to 'higher order hypercomplex cells' (Hubel & Wiesel, 1962).

Prosopagnosia can arise as a result of the damage at any one of these stages or phases (Bauer, 1986; DeRenzi, 1986; Damasio, Damasio, & Tranel, 1986). Some people may not be able to recognize faces as faces, while others cannot recognize familiar faces (the most traditional notion of prosopagnosia), and still others can admit that the person seems familiar, yet strangely enough, they do not know anything about them (Takamura, 1996). All of the prosopagnosians seem to be unable to integrate partially high complex featural stimuli to an overall binded face.

However, how is the computational problem of fast and accurate binding ever possible to be realized? With his *temporal correlation* hypothesis, von der Malsburg (1981) proposed an appealing coding mechanism that would solve the feature integration or binding problem in sensory processing. The key idea is that cells express their participation in the representation of the same external object by synchronized firing (see also Kappen, 1997), or cell assemblies show oscillatory responses, respectively (Engel, Konig, Kreiter, Schillen, & Singer, 1992; Gray, König, Engel, & Singer, 1989; Singer & Gray, 1995). This means that the essential part of a complex cognitive process is not found in single cells, but in the concerted activity of a population of neurons distributed over the cortical surface (Elliott & Müller, 1998); this argumentation is especially relevant for the recognition of faces. If all such groups of neurons and their combined firing were destroyed, then a person with such an assumed damage would not see a face, but only parts of a face, such as the eyes, the nose, the mouth, etc. (Crick & Koch, 1998).

[30] Kant already pointed out, that the temporal order and binding is essential for formulating pure philosophical thoughts. He called this binding feature "the transcendental unity of apperception" (Kant, 1968).