How We Detect a Face: A Survey of Psychological Evidence

Michael B. Lewis, Hadyn D. Ellis
School of Psychology, Cardiff University, PO Box 901, Cardiff, CF10 3YG, United Kingdom

ABSTRACT: Scientists strive to build systems than can detect and recognize faces. One such system already exists: the human brain. How this system operates, however, is far from being fully understood. In this article, we review the psychological evidence regarding the process of face detection in particular. Evidence is collated from a variety of face-processing tasks including stimulus detection; face categorization, visual search, first saccade analysis, and face detection itself. Together, the evidence points towards a multistage-processing model of face detection. These stages involve preattentive processing, template fitting, and template evaluation. Comparisons are made with automatic face-detection systems. © 2003 Wiley Periodicals, Inc.

Key words: face detection; face recognition; visual array; luminance

1. INTRODUCTION

Great advances have been made in the automatic analysis and synthesis of facial images. Many of these advances are described in this issue. Here, we would like to describe what is known about one specific facial-image-processing system that, in many ways, outperforms even the most sophisticated automatic systems. The human brain is able to detect and recognize a target face from an extensive gallery in well under a second. The process occurs accurately and apparently effortlessly. Much is known about the psychology of how a face is recognized and reviews of this extensive literature are available (Valentine, 2000). Only a limited amount of research, however, has focused on the processes involved in face detection by humans. Here, it is our intention to review this evidence and relate it to the more extensive research into automatic-face-detection systems.

The detection of a face in a visual scene is a prerequisite to any further processing, such as identification or expression analysis. These subsequent tasks have received a great deal of attention by psychologists and, consequently, much is known about how we recognize faces and emotions. This information comes from analyses of the factors that affect efficient processing of faces. It was found, for example, that inverting a face greatly hinders identification (Yin, 1969) and emotional encoding (McKelvie, 1996). Presenting a face in a negative luminance pattern, however, affects identification to a lesser degree and has no effect at all on emotion recognition (White, 2001).

In studies such as these, reaction times can be just as informative as accuracy scores. It can be assumed that if a factor increases reaction times then it is important in that process (Donders, 1868/1969). Further, interactions between factors can distinguish whether the two factors affect the same stage of processing or not (Sternberg, 1969)—put simply, if the factors interact then they affect the same stage (or parallel stages), whereas if the factors are additive then they affect different (serial) stages. This type of analysis of reaction-time data provides us with the tools required to investigate face processing. Recently, these same tools have been brought to bear on the issue of how we first detect faces in a visual scene. Here, we will review what we currently know about face detection by humans from these studies. Comparisons will be drawn between how humans deal with this problem and how some automatic systems deal with the same problem.

First, we will consider whether face detection is any different from the detection of other objects. This is followed by a discussion of four bodies of research into face detection by humans, each of which provides an insight into the perceptual processes. The first three areas of research were not specifically designed to explore face detection in the natural world but provide a useful beginning for discussion. The last area, and most recent research, is a concerted attempt to identify how humans detect faces in the visual world.

2. IS FACE DETECTION SPECIAL?

Face recognition is considered by some to be special and different from the recognition of other stimuli (Ellis et al., 1989). Evidence for this comes from: behavioral measures of inverted or jumbled faces (Yin, 1969; Tanaka and Farah, 1993); developmental sources (Johnson et al., 1991); brain imaging sources (Aguirre and Farah, 1999); and neurological sources (Bauer, 1984). Studies of expertise in other aspects of recognition, however, raise the issue of whether the special nature of face recognition, in fact, is due to our expertise with such images (Diamond and Carey, 1986).

Face detection, like recognition, may also be special and distinct from the recognition or detection of other classes of stimuli. Faces cannot easily be broken down into geons (3D primitives of objects such as cylinders or cones) as has been proposed for the discrimination of other types of objects (Biederman, 1987). A head is just a spherical shape with a few indentations and lumps. Research by Schyns, however, has identified similarities in the processes of face detection and object recognition (Schyns, 1998). Indeed, it may just be our expertise with faces and their social importance that makes their detection special. Understanding face detection, therefore, could reveal a great deal about object detection and recognition in general as well as informing us about how one can perform this important element of pattern recognition.
3. THE FACE-DETECTION EFFECT

Reaction times are not the only performance measure used in perceptual studies. Another useful measure is this minimum presentation time required for a stimulus to be detected or the percentage detection at a particular presentation time. In such cases, it is typically the presence or absence of any image that must be detected. The task does not involve discrimination of the image and so it is not, therefore, directly comparable with the task of face detection in its usual sense. Purcell and Stewart (1986, 1988) are pioneers of this methodology, and they discovered what they called the face-detection effect.

Purcell and Stewart (1986) found that the presentation time required to detect a stimulus if it was a face was shorter (38 ms) than the time required to detect a stimulus if it was a jumbled face (56 ms). These times varied when using different masking procedures (stimuli immediately before and after the stimulus), but there was consistently an advantage for the intact faces. Other studies have found a similar advantage for upright faces over inverted faces (Purcell and Stewart, 1988), and the same effect appears using line-drawn faces or when faces are presented away from the centre of attention.

The face-detection effect shows that, early on in the perceptual process, the detection of the mere presence of a stimulus is influenced by its structure. If the structure contains a face then it is more likely to be detected than if it contains a visually equivalent nonface image (a jumbled or inverted face). This means that the face is a special object for us in terms of its detection. It suggests the possibility that there is a dedicated system for the detection of faces, which is not involved in the detection of other objects or, at least, there is a general system that is highly specialized toward detecting faces. Without doubt, we are better able to detect an object if it is a face than if it is not. The experiments conducted into the face-detection effect, however, do not tell us why faces are special or how they are detected.

4. FACE CATEGORIZATION

Useful insights into the detection of faces come from studies of face-categorization tasks. In these tasks, subjects are presented with a face or a jumbled face and they have to make a speeded response as to whether the face is jumbled or not. Figure 1 shows the type of stimuli used in such a task. This can be considered to be a face-detection task because a positive response is required whenever the stimulus is an intact face. The task, however, is not equivalent to natural face detection in the real world for two reasons. First, the stimulus image is always presented to the center of the visual field with no surrounding context. Second, the images that act as distractors are unnatural and contain many face elements. A jumbled face still contains most of the elements of a face and only the arrangement is different. The task in this experiment, therefore, requires interpretation of their relative positions of features rather than feature or face detection. The face-categorization task is not an ideal task for finding factors that affect face detection, but it was not designed to be. The data that this task produces, however, may provide some limited insight into the task of detecting a face.

The face-categorization task, developed by Valentine and Bruce (1986), has been used to explore whether a face’s rated distinctiveness determines the speed of the categorization decision. Distinctiveness ratings are derived from asking people how easy a person would be to recognize. These ratings are consistent across subjects and correlate with accuracy and speed of recognition [distinctive faces are recognized faster than typical faces (Valentine and Bruce, 1986)] and attractiveness [distinctive faces are less attractive (Vokey and Read, 1992)]. Accounts for these distinctiveness effects have involved a multidimensional face-space metaphor (much like the space spanned by eigenfaces) in which faces are encoded. Distinctive faces are a long way from the center of this space and have few near neighbors. Recognition is easier for distinctive faces because fewer other known faces are competing for recognition.

The effect of rated distinctiveness was investigated in a face-categorization task reported by Valentine and Bruce (1986). Unfamiliar faces were presented in greyscale. These faces were either normal or were jumbled. To form a jumbled face, equal rectangles containing the eyes, the nose and the mouth were cut out and interchanged. The subjects’ task was to indicate if each image was an intact face or a jumbled face. The effect of distinctiveness on this task was found to be the reverse of the effect on the recognition task. Distinctive-rated faces were categorized more slowly than typical-rated faces.

The typical-face advantage in the face-categorization task was accounted for by a face-schema explanation, which later gave way to the multidimensional face-space account. In fact, the face-space account included two different explanations for the effects (Valentine, 1991). The first, the norm-based account, is that we have a single concept of an archetypal face against which to compare any new image. The closer the stimulus is to this archetypal face (or norm face or prototypical face) then the easier and more confident we are to say that an image is a face. The second, the exemplar-based account, suggests that the image is simultaneously compared against all known faces. A sum-of-similarity algorithm is then used to produce the confidence that the image is a face. A face that is like many known faces (i.e., a typical face) will be categorized faster than a face that is like just a few known faces (i.e., a distinctive face).

The norm-based account of the face-categorization task can be compared to a deformable-template matching procedure. The norm face is the template against which the stimulus is compared. The closer the image is to the template the faster the face-categorization decision is made. It can be argued that the exemplar-based account is identical to the norm-based account because of a lack of specificity. The sum-of-similarity algorithm is not defined and, with an appropriate algorithm, the exemplar-based account would be the same as the norm-based account.

Further investigation into the face-categorization task explored the effect of inversion and how it interacts with distinctiveness (Valentine, 1991). Inverting the image made face-categorization
slower. Moreover, the effect of distinctiveness did not interact with the effect of inversion (i.e., the distinctiveness effect was approximately the same for upright and inverted faces). Using what we know about reaction times, we can infer that inversion and distinctiveness affect different aspects of the face-categorization task. It was just mentioned how distinctiveness may reflect the degree of similarity between a face template and the image. If a template-matching procedure were used on an inverted face, then obviously it would give a very poor correlation. Instead, the template may have to be rotated (or the mental image may have to be rotated) in order to begin the template-matching procedure. Alternatively, we may have two templates: one for upright faces and one, less efficient, for inverted faces.

The face-categorization task has also been used to assess neurological patients who have the condition of prosopagnosia (selective impairment of face recognition). It has been found that some prosopagnosics can detect faces but others cannot (de Gelder and Rouw, 2000). This suggests that the systems responsible for face recognition and face detection (or at least discrimination from a jumbled face) can be differentially damaged and so are probably located at different places in the brain.

Although it is too early to be able to say what the template for the face-categorization task is, it is possible to make some statements about its origin. Studies using faces from other races suggest that the template is learnt from the faces encountered. Other-race faces produce longer reaction times in this task. A purely learnt template, however, cannot be the whole story. Evidence from neonates (less than an hour old) found an advantage for tracking a face-like stimulus over a jumbled stimulus (Johnson et al., 1991). It would appear, therefore, that there is some innate template that we develop over time that becomes tuned to faces of a particular race through experience.

5. VISUAL SEARCH

The face-categorization studies mentioned in the last section are not strictly face-detection experiments because the faces are presented on their own and the distractors contain facial features. A series of experiments addressed this first problem by using a visual search paradigm. The visual-search paradigm was developed by Treisman and Gelade (1980) as a method for distinguishing between serial and parallel visual processing. The basic experiment involves detecting a single item in a set of distractors, for example, a red circle among many yellow squares. In this example, the red circle immediately "pops out" of the image, and if the time taken to detect it is unaffected by the number of distractors, the search is deemed to be parallel. If, however, the task is to detect a red circle among red squares and yellow circles, then the target item does not pop out. The search requires more effort, and the time taken to do the task increases as the number of distractors increases. This type of search is deemed to be serial.

Wolfe (1994) updated these concepts of parallel and serial search and introduced the idea of preattentive guided search. Items that pop out do so because there is some element that can be located preattentively (and therefore in parallel) and allows attention to be focused immediately on that item. This interpretation still means that the visual-search paradigm is a useful method for investigating the processes involved in face detection.

There have been several studies that have applied the visual-search paradigm to faces. Some of these explored issues such as race effects and so are not of importance here (Levin, 2000). Nothdurft (1993) explored whether faces (and facial expressions) pop out in a visual-search paradigm. In one of his experiments, subjects were presented with one line-drawn schematic face in an array of jumbled faces of the same kind. Half of the arrays contained a face, half did not, and the subjects task was to indicate whether a face was present or not. Array sizes varied up to 48 items. Figure 2 shows the type of array used in this task. The relationship between array size and positive reaction times was found to be 113 ms per item. The size of this gradient leads to the conclusion that faces do not pop out, and so face detection in this situation is serial. Similar studies were conducted by Kuehn and Jolicoeur (1994), but their schematic faces included basic shading of skin and hair tones. The images were still obviously drawn faces but they did more closely resemble real faces than Nothdurft's stimuli. This later study also used much smaller arrays ranging from two to four items. The gradient of the reaction-time-array-size slope in these experiments was approximately 50 ms per item. This slope, however, appears to be affected by the nature of the distractors such that the less face-like the distractors, the more shallow the slope is.

One major limitation of Nothdurft's (1993) and Kuehn and Jolicoeur's (1994) studies was the use of abstract schematic faces. These are unrealistic stimuli and so it is difficult to generalise the data to real face detection. Brown et al. (1997) conducted a similar visual-search study but used real faces and jumbled faces. Rather than using reaction times, this study used eye tracking to detect where the subject looked first after presentation of an array of three jumbled faces and an intact face. It was found that, with naive subjects, the first saccade (movement of the eye) did not go toward the face with anything above chance probability. For practiced subjects, there was a speed-accuracy trade-off such that they were above chance but the time until the first saccade was longer. The conclusion from this study is that a face cannot be distinguished easily in peripheral vision from a jumbled face. Focused attention, therefore, is required in order to detect faces but only in the limited condition of where the distracting images are jumbled faces. In natural face detection, the context or scene is unlikely to be made up of jumbled faces. How we conduct face detection in such a situation, however, may be different from that in the situation described above. The last set of studies we will describe looked specifically at the detection of faces in naturalistic scenes.

6. NATURALISTIC FACE DETECTION

Lewis and Edmonds (2002) conducted a series of experiments with the aim of revealing important factors affecting naturalistic face
detection. The first of these experiments leads on from the section above because it used the visual-search paradigm. A visual array was generated using random rectangles cut from a natural visual scene (captured from TV stills). In half of these arrays, one of the rectangles contained a face. The arrays were made up of 4, 9, or 16 rectangles and were presented in color. The subjects’ task was to detect the presence or absence of a face in each array. Reaction times were recorded and a reaction-time-array-size slope was obtained. When the face was present, the slope had a gradient of just 3 ms per item—a level that is indicative of a face popping out of the array. A conclusion from this study, therefore, is that for color photographic faces in a naturalistic (albeit jumbled) scene, face detection does involve preattentive processing. Some feature of the face can be found using automatic and parallel analysis of the image, and this feature is used to guide attention. What this feature is remains to be seen but one strong possibility is that it is the color of the face. This result is contrary to previous research and, therefore, demonstrates the importance of using naturalistic scenes in the analysis of face detection.

The experiment just described explored face detection using naturalistic distractor cells (parts of the original image that not containing the face), but these cells were arranged in a random array. In a natural scene, we do not see a face in isolation or surrounded by random artefacts. It is most likely that we see a face above a body and pair of shoulders. The context in which a face appears may affect how easily it can be detected. Information from the scene may act to guide detection if that scene contains appropriate cues. Lewis and Edmonds (in press) conducted an experiment that explored the role played by the scenic context in which a face appears. The speed of face detection was compared between faces that appeared in intact natural scenes and those that appeared in jumbled scenes (gridlines were overlaid on both types of images to create standard images). Figure 3 shows examples of the stimuli used.

Results showed that subjects were consistently faster at detecting a face when it appeared in an intact scene than when it appeared in a jumbled scene. The size of this scene advantage was approximately 60 ms. This result means that the process of searching for candidate locations for faces can be guided by the contextual, nonfacial information present in the scene. The task of face detection, therefore, involves much more than just finding a face but also includes analysis of the entire visual image for clues about location. Context and, possibly, color appear to guide attention toward candidate face locations. These clues, however, are of little use in arriving at a decision that a region is actually a face. In order to investigate the face detection procedure, experiments were conducted that presented either just faces or parts of naturalistic scenes (Lewis and Edmonds, in press). These images were transformed in a number of ways, and the effects the transformations had on reaction times to a face-detection task were analyzed. The transformations were inversion (180° rotation in the plane), luminance reversal (reversing the difference of the luminance of each pixel from the average luminance), hue reversal (based on 180° change on a rough color wheel), reduction of contrast by 50%, and a small degree of blurring.

Luminance reversal had the greatest effect on the speed to detect the isolated face (35 ms). Inversion also had a large effect (16 ms), whereas the reversal of hue had very little effect at all (8 ms). This last point suggests that, once a candidate region has been attended to, color plays little role in face detection. The inversion effect did not interact with any of the other four transformations. This suggests that inversion affects a different stage of processing to the other transformations. As one would expect, blurring and contrast reduction each increased the time required to make a face-detection decision. Further, these two factors interacted such that the detriment was greater for a blurred, contrast-reduced face than would be expected by adding the detriment for each transformation on its own. This suggests that these two transforms affect the same stage of processing. Because the interaction with luminance reversal was not tested, it is possible that this stage is the same as is affected by this transformation.

Results from these experiments into focused face-detection reveal that at least two further stages are required. One stage is only affected by inversion (of the transformation tested), whereas the other is affected by contrast reduction, blurring, and luminance reversal. The former of these two stages could be interpreted as a template-matching procedure in which a deformable template of a face must be fitted to the presented stimuli. Obviously, the more common upright matching would be conducted first or faster than matching an unusual inverted face. Once the template has been deformed to best fit the pattern within the candidate region, a comparison is made between the luminance pattern and the image and that expected by template. The correlation between the luminance pattern of the image and the template will be greatly reduced for luminance reversed images (indeed, it will be negative). For such images, an edge-detection template may be used. Contrast reduction and blurring would also reduce the degree of similarity between template and image leading to the observed interaction. This may also be the stage at which the distinctiveness of the face, as observed in the face-categorization tasks (Valentine and Bruce, 1986), has its effect—particularly if the template is based on an average face.

7. A MODEL OF FACE DETECTION

Although the data from psychological enquiry into face detection are at an early stage, they do allow a model to be proposed. This model is intended to be a starting point for further research and is not meant to be a definitive account. The model is based on the current data available and is hoped will provide guidance for future studies into this area.

The model has three stages. The first stage is a parallel, preattentive scanning of the visual scene. This stage picks up global contextual cues and local face-unique cues (possibly color). These cues help to identify candidate regions that may contain a face.
Attention is focused upon this region, and the second stage begins. This stage involves the deformation of a face template through skew and rotation (in the plane or out of the plane) to best fit the candidate region in terms of the general pattern of edges. The third stage involves evaluating the degree of fit between the luminance pattern in the candidate region and the deformed template. The better the fit between these, the faster the response will be.

This basic model can be tested, and possibly falsified, in many ways through planned experimental work. Factors affecting each stage can be predicted and tested to see if they interact with other factors that are thought to affect those stages. For example, it can be predicted that rotation or skewing a face will make the second stage slower but such transformations should also interact with inversion. Such predictions are eminently testable.

8. SUMMARY AND CONCLUSIONS

The proposed model of face detection by humans may have some similarities to methods used by computer scientists in their efforts to build automatic face-detection systems. Such systems were extensively reviewed by Yang et al. (2001). From such a review, it is possible to pick out similarities. For example, color has been used extensively to identify face candidate regions by some systems (Lee et al., 1996; Wu et al., 1999). In addition, edge detection has been used to fit face templates (Han et al., 2000). Further, matching to template is a common practice in automatic face detection since the introduction of eigenfaces (Turk and Pentland, 1991). There is, therefore, a high degree of similarity between how humans deal with the problem of face detection and how computers do it. One difference, however, is that humans use information from the visual context to guide detection. To our knowledge, no automatic system uses such a method.

REFERENCES


S.M. Kuehn, and P. Jolicoeur, Impact of quality of the image, orientation and similarity of the stimuli on visual search for faces, Perception 23 (1994), 95–122.


