ULTRA RAPID FACE DETECTION IN NATURAL IMAGES : IMPLICATIONS FOR COMPUTATION IN THE VISUAL SYSTEM

Simon J. Thorpe, Sébastien Crouzet, Holle Kirchner and Michèle Fabre-Thorpe
Centre de Recherche Cerveau & Cognition, UMR 5549
Université Paul Sabatier, Toulouse, France
simon.thorpe, sebastien.crouzet, holle.kirchner, michele.fabre-thorpe @cerco.ups-tlse.fr

ABSTRACT
Using a choice saccade task, Kirchner and Thorpe recently demonstrated that detection of animals in natural scenes is considerably faster than previously supposed [1]. Here we present some new data with the same task that show that face detection is even more efficient. When two images are flashed to the left and right of fixation, subjects can accurately make saccades to the side where there is a human face with a mean reaction of time of only 147 ms and an accuracy level of 94%. The earliest reliable saccades were made as early as 110 ms after stimulus onset. If we allow roughly 20 ms for saccade initiation, such data leaves very little time for visual processing and seems to rule out computations that require more than one spike per neuron. Furthermore, it seems clear that only a feed-forward pass through the visual pathways can be performed in so little time.

KEY WORDS
Visual Processing, Feed-forward Processing, Natural Images, Neural Coding

1. Introduction

There is considerable debate about the relative roles of feed-forward and feed-back pathways in sensory processing. While everyone would agree that both are used, it is important to ask whether certain types of processing can be performed using just the feed-forward pass. One way to address this issue is examine the sort of visual processing that can be done very rapidly. About a decade ago, we demonstrated that humans are very good at deciding whether a briefly flashed natural scene contains an animal [2]. Reaction times in a go/nogo task were around 400 ms, with the earliest reliable responses occurring at around 280 ms. Furthermore, simultaneously recorded Event Related Potentials (ERPs) showed dramatic differences between target and distractor trials from about 150 ms, and this value has often been used as an estimate of the time required for visual processing. Nevertheless, using a new choice saccade task, Kirchner and Thorpe recently showed that when two images are briefly flashed to the left and right of fixation, subjects can make saccades to the side where there is an animal with reaction times as short as 120 ms [1]. Mean accuracy was 90% and mean reaction times were 228 ms, a lot faster than in the more standard manual go/nogo task.

Previous work had shown that in the go/nogo task, subjects are also very good at responding to photographs of human or animal faces [3]. When the target category was a human face, subjects performed at 97.7% correct, but they performed just as well when the target category was an animal face (97.9%), despite the fact that the very same images were being used in both conditions. This ability to switch target category at will is strong evidence for task related modulation of processing at some stage of the decision process.

Given this remarkably high level of performance, we were interested to know how well subjects would perform in a choice saccade version of the task.

2. The Choice Saccade Task

The experiments were performed using a protocol that was very similar to the one used by Kirchner and Thorpe with the exception that the images were left on for 400 ms rather than simply flashed for 20 ms (Figure 1). It seems that this variation allows the saccade initiation process to be performed more reliably, and although it does not affect early latencies it reduces variance and associated mean reaction times.

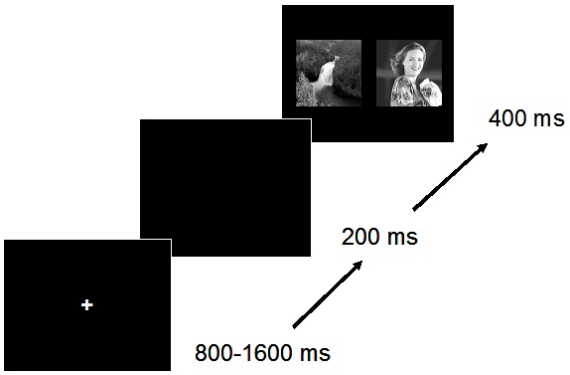

Figure 1 : Basic stimulus configuration used in the choice saccade task. A fixation cross is shown for a variable period followed by a 200 ms period with no stimulus. Then, two images are presented for 400 ms to the left and right of fixation. The subject's eye movements are recorded using the electro-oculogram (EOG) and trials classified as correct or incorrect depending on the direction of the initial saccade.

Although a large number of different variations on this basic paradigm have now been used, we will here concentrate on just one set of experiments that examined the performance of human subjects with human faces as targets. The target images were a varied set of 250 human faces at different scales and the distractors a very wide range of natural images containing buildings, plants, signs etc..

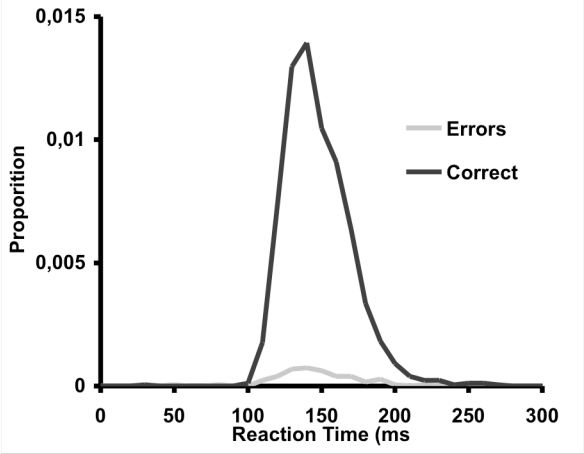

Figure 2 : Reaction time distributions for human face vs distractors in the choice saccade task. The curves plot the results for 6 subjects.

The basic results are illustrated in Figure 2 that shows the reaction time distributions for saccade onset in six subjects. The mean RT was just 147 ms, ranging from 131 ms for the fastest subject to 164 ms for the slowest, but in all cases, accuracy was high, ranging from 93.5% to 96.6% with an average across the subjects of 94.4%. Remarkably, this high level of accuracy was seen for the very fastest responses, namely for those initiated between 100 and 109 ms after stimulus onset. For one of the best subjects, the result was even more striking because every one of the 52 saccades that the subject made with onset latencies of 123 ms or less was in the correct direction – i.e. in the direction of the face. The probability of obtaining such a result by chance is infinitesimally small. Figure 3 shows some examples of the sorts of image pairs that could be successfully processed with these very short reaction times. What does the ability of subjects to perform this task so well and so fast tell us about the underlying neural computations? In the next section, we will discuss why this result has such profound implications.

3. Brain Mechanisms for Face Processing

It has been known for a long time that the primate brain contains neurons selective for faces [4]. These neurons can fire at latencies of 80-100 ms in monkeys, which in principle would seem to be sufficiently rapid to allow the triggering of saccadic responses in humans from around 100 ms onwards if we allow roughly 20 ms or so for the initiation of the eye movement itself. However, this ignores the fact that humans are known to be at least 50% slower than macaques on a wide range of tasks [5]This has been termed the 3/5 rule [6]. For example, in the go/nogo manual categorisation tasks used extensively in our group, human reaction times are almost never shorter than about 270 ms, whereas monkeys can perform the exact same task with behavioural reaction times that start at 160-180 ms [7]. One plausible explanation for this reaction time advantage for macaques could be the difference in head size between humans and monkeys [5]. Experimental evidence indicates that intra-cortical conduction velocities may be as slow as 1-2 m/s, which means that much of the increase in latency seen as one steps through the ventral visual pathway from V1 to IT via V2, V4 and Posterior IT could in fact be due to conduction delays. But, if true, this would imply that the equivalent latencies for face selective mechanisms in humans would be expected to be considerably longer than would be needed to trigger behavioural responses fast enough for the current data.

Figure 3 : 4 pairs of images that were used in the choice saccade task. All four pairs resulted in saccades towards the side with the face in less than 120 ms.

Interestingly, in monkeys, recent neurophysiological evidence has demonstrated that the latency of single unit responses to human and monkey faces is substantially shorter than to faces of other animals, even when the absolute size of the response was matched [8].

In humans, there is a very extensive literature on ERPs that are thought to be specific to faces. In particular, many groups have investigated the N170 potential which peaks at around 170 ms, but actually starts considerably earlier (see for example, [9]). More recent work using Magnetoencephalography (MEG) has found evidence for a face-selective response with a latency of around 100 ms (the M100) [10]. There have even been reports of face selectivity in ERP responses that start considerably before 100 ms [11], although it has been unclear what the functional significance of these effects could be.

However, in the light of the current behavioural data, it seems clear that these very rapid selective responses to face stimuli could indeed be important. But this raises the issue of how these selective responses can be produced so rapidly

4. Possible Computational Mechanisms

For a number of years we have argued that the speed of processing poses a serious problem for the widely accepted view that the underlying processing can use the firing rate of individual neurons or even populations of neurons. The problem is that when the observation period is 10 ms or less, few neurons will have time to generate more than one spike, thus making it impossible to estimate instantaneous firing rate with any precision [12, 13]. An alternative strategy would be to use the fact that the most strongly activated neurons will tend to fire first. As a result, much can be learned by simply monitoring the order of firing across a population of cells. Even without having a specific reference, knowing that a particular group of cells fired first in response to a flashed stimulus will provide a great deal of information about the stimulus. For example, using a very simplified model of the retina, VanRullen and Thorpe noted that even when less than 1% of the retinal ganglion cells have fired a single spike, it is possible to reconstruct the incoming image sufficiently well to identify most objects in natural images, at least as long as there is one reasonably large object in the scene [13].

More recently, Guyonneau and coworkers looked at how Spike Time Dependent Plasticity (STDP) affects the distribution of synaptic weights in the afferents to a neuron that is repeatedly stimulated with temporally structured waves of spikes. The first point is that STDP systematically leads to high weights being concentrated on the inputs that fire earliest [14]. Importantly, this is true irrespective of the actual firing rates of the afferents. In the end, it is only those afferents that reliably fire at short latency that can "win" in the competition to get the highest synaptic strengths. This phenomenon has another important consequence. At the beginning of training, the response latency of a given neuron to a visual input will always be relatively long, because initially none of the afferents have particularly high connection strengths and so a large number of them need to be activated before the post-synaptic neuron can reach threshold. However, with repeated presentation of the same visual stimulus, the neuron will progressively fire earlier and earlier as different sets of afferents are strengthened and then depressed. In the end, only a few very strong synapses are needed to push the neuron over threshold, and these are the afferents that happen to fire reliably and quickly in response to the stimulus [15].

In the latest set of studies using STDP, Timothée Masquelier has demonstrated that when the same sorts of ideas are applied in the case of a hierarchically organised visual system, the higher levels of the system will naturally develop selectivities that correspond to frequently occurring visual patterns in the environment [16]. He showed that this was the case for face stimuli by presenting the system images from the CalTech face database, a large set of images of faces seen under various lighting conditions and with very varied backgrounds. After a few hundred presentations, neurons in the recognition layer begin to show selectivity for face-like patterns, and over several thousand presentations, the neuronal responses become both more and more selective and increasingly rapid.

Could it be that this sort of prolonged experience dependent learning is involved in generating the exceptionally fast behavioural responses described here? It is certainly true that faces are very important stimuli for humans, and it is likely that right from birth, human infants will preferentially look at visual patterns that are sufficiently close to faces.

Figure 4 : Simple visual patterns composed of just four lines that produce activation of "face" level representations.

These findings raise the following possibility. Perhaps processing for faces is so over-trained that the neurons in higher order areas can detect the presence of a combination of features that is indicative of the presence of a face even when very little information is available. To take an extreme case, consider the very simple stimuli illustrated in Figure 4. Despite the fact that each is composed of only 4 lines and could in principle be coded by the simultaneous activation of just four orientation selective neurons in cortical area V1, they are nevertheless sufficiently rich to evoke the concept of "face" in the vast majority of observers. Indeed, it is remarkable how with just a very few contours it is possible to generate drawings that are capable of activating quite complex representations (this is the principle of the game "Pictionary" where players attempt to evoke an idea using a minimum number of lines).

Our suggestion is that the combination of a wave of spikes with STDP means that neurons in higher order visual areas will naturally become selective to

frequently occurring visual stimuli, and that over time, they become tuned to the particular combinations of visual elements that correspond to the earliest firing neurons in areas such as V1. Such a mechanism could go a long way to explaining how human observers can make such rapid responses in our choice saccade task.

5. Conclusion

The present paper gives just a sketch of the sort of mechanism that will be needed to explain the phenomenal speed with which humans can perform the choice saccade task with faces as targets. But it raises a number of important questions that currently do not have clear responses. For example, it appears that only a few milliseconds of activity in areas like V1 may be enough to trigger face selective responses in later areas. How many neurons will actually fire during the first 5-10 ms of processing of a natural image in an area like V1? The fact is that currently we have no answer to such questions. Much needs to be learned, but it seems likely that the temporal constraints imposed by our recent experiments using the choice saccade task will provide a critical test for many models.

Acknowledgements

The authors thank the ANR (NatSTATS), the European Commission (Decisions in Motion) and the CNRS for support.

References

[1] H. Kirchner and S. J. Thorpe, *Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited*, *Vision Res, 46*, 2006, 1762-76.

[2] S. Thorpe, D. Fize and C. Marlot, *Speed of processing in the human visual system*, *Nature, 381*, 1996, 520-2.

[3] G. A. Rousselet, M. J. Mace and M. Fabre-Thorpe, *Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes*, *J Vis, 3*, 2003, 440-55.

[4] D. I. Perrett, E. T. Rolls and W. Caan, *Visual neurones responsive to faces in the monkey temporal cortex*, *Exp Brain Res, 47*, 1982, 329-42.

[5] M. J. Mace, G. Richard, A. Delorme and M. Fabre-Thorpe, *Rapid categorization of natural scenes in monkeys: target predictability and processing speed*, *Neuroreport, 16*, 2005, 349-54.

[6] J. J. Foxe and C. E. Schroeder, *The case for feedforward multisensory convergence during early cortical processing*, *Neuroreport, 16*, 2005, 419-423.

[7] M. Fabre-Thorpe, G. Richard and S. J. Thorpe, *Rapid categorization of natural images by rhesus monkeys*, *Neuroreport, 9*, 1998, 303-8.

[8] R. Kiani, H. Esteky and K. Tanaka, *Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces*, *J Neurophysiol, 94*, 2005, 1587-96.

[9] S. Bentin, T. Allison, A. Puce, E. Perez and G. Mccarthy, *Electrophysiological Studies of Face Perception in Humans*, *J Cogn Neurosci, 8*, 1996, 551-565.

[10] J. Liu, A. Harris and N. Kanwisher, *Stages of processing in face perception: an MEG study*, *Nat Neurosci, 5*, 2002, 910-6.

[11] Y. Mouchetant-Rostaing, M. H. Giard, C. Delpuech, J. F. Echallier and J. Pernier, *Early signs of visual categorization for biological and non-biological stimuli in humans*, *Neuroreport, 11*, 2000, 2521-5.

[12] J. Gautrais and S. Thorpe, *Rate coding versus temporal order coding: a theoretical approach*, *Biosystems, 48*, 1998, 57-65.

[13] R. VanRullen and S. J. Thorpe, *Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex*, *Neural Comput, 13*, 2001, 1255-83.

[14] R. Guyonneau, R. Vanrullen and S. J. Thorpe, *Neurons Tune to the Earliest Spikes Through STDP*, *Neural Comput, 17*, 2005, 859-79.

[15] R. Guyonneau, R. Vanrullen and S. J. Thorpe, *Temporal codes and sparse representations: A key to understanding rapid processing in the visual system*, *J Physiol Paris, 98*, 2004, 487-97.

[16] T. Masquelier and S. J. Thorpe, *Unsupervised learning of intermediate complexity visual features through Spike Timing Dependent Plasticity*, *Neuro Comp 2006*, 2006.