Biosystems, 1998, 48(1-3), 229-239.

FACE PROCESSING USING ONE SPIKE PER NEURONE.

Rufin Van Rullen, Jacques Gautrais, Arnaud Delorme and Simon Thorpe
Centre de Recherche Cerveau & Cognition, UMR 5549,
133 route de Narbonne,
31062 Toulouse, France.

Abstract

The speed with which neurones in the monkey temporal lobe can respond selectively to the presence of a face implies that processing may be possible using only one spike per neurone, a finding that is problematic for conventional rate coding models that need at least two spikes to estimate interspike interval. One way of avoiding this problem uses the fact that integrate-and-fire neurones will tend to fire at different times, with the most strongly activated neurones firing first (Thorpe, 1990). Under such conditions, processing can be performed by using the order in which cells in a particular layer fire as a code. To test this idea, we have explored a range of architectures using SpikeNET (Thorpe and Gautrais, 1997), a simulator designed for modelling large populations of integrate-and-fire neurones. One such network used a simple four-layer feed-forward architecture to detect and localise the presence of human faces in natural images. Performance of the model was tested with a large range of grey-scale images of faces and other objects and was found to be remarkably good by comparison with more classic image processing techniques. The most remarkable feature of these results is that they were obtained using a purely feed-forward neural network in which none of the neurones fired more than one spike (thus ruling out conventional rate coding mechanisms). It thus appears that the combination of asynchronous spike propagation and rank order coding may provide an important key to understanding how the nervous system can achieve such a huge amount of processing in so little time.

1. Introduction.

Electrophysiological data indicate that some neurones of the monkey temporal lobe respond selectively to complex stimuli such as faces with a latency of 80-100 ms after stimulus onset (Bruce et al.,1981; Perrett et al., 1982; Oram & Perrett, 1992 Jeffreys, 1996). Within this short time, information has to be processed not only by the retina and LGN, but also by several cortical areas, including V1, V2, V4 and PIT, in each of which 2 synaptic stages at least must be passed through. Hence information has to run through 10 different processing stages, within the 100 ms-window considered.

Furthermore, conduction velocities of intracortical fibres are known (Nowak & Bullier, 1997) to be remarkably slow (e.g. < 1 m/s), which leaves less than 10 ms for computation (axonal conduction, synaptic transmission, somatic integration, and spike emission) in each processing stage. Such rapid processing poses severe problems for conventional rate coding mechanisms, since very few cells will be able to emit more than one spike in this time, and although it would be possible to calculate firing rates across a population of cells, we have argued elsewhere that this would require very large numbers of redundant cells (Gautrais & Thorpe, 1998).

How might the visual system perform complex tasks like face detection without emitting more than one spike in each processing stage? One option is to use differences between spike latencies across a population of neurones as a code (Thorpe, 1990). This argument is based on the fact that integrate-and-fire neurones, such as those observed in the visual cortex, will tend to fire at different times, with the most strongly activated neurones firing first (fig. 1). Only one spike per neurone is then required to create a complete representation of the input stimulus.

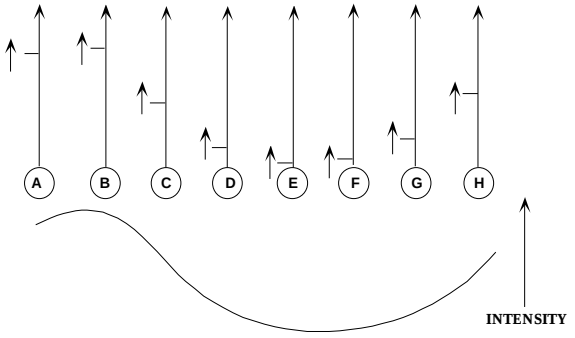

Figure 1. Neurones can act as analogue-latency converters, with more strongly activated neurones firing first. One can also use the order of firing (B > A > H > C > G > D > F > E) as a code. With 8 neurones, there are 8! i.e. 40 320 different possible orderings.

Under such conditions, one can also use the relative order of firing of units in a particular layer to represent the input information. This sort of coding scheme, which we call Rank Order Coding, is still compatible with the constraint of having only one spike per neurone, whereas it greatly simplifies the computation. Nevertheless, it does not lead to a decrease in representational capacity by comparison with traditional rate coding, since a population of n neurones can actually discriminate n! different stimuli, when in the same time window only n+1 codes could be recognised with rate coding (see Gautrais & Thorpe, 1998, for a theoretical analysis of these issues).

In order to test this new coding scheme, a simulator called SpikeNET was designed, which allows to model very large populations of integrate-and-fire neurones (Thorpe & Gautrais, 1997). We used this simulator to evaluate the performance of a simple model based on Rank Order Coding, and to test whether complex visual processing tasks such as face detection and localisation in a natural image could be performed on the basis of only one spike per neurone.

2. Architecture of the model.

2.1. Asynchronous propagation.

SpikeNET simulates neurones with simple integrate-and-fire characteristics: afferent spikes increase their activation level, until they reach a threshold and fire a single spike. The response of such a neurone (i.e. the latency of its output spike) can be made to depend upon the relative order of firing of its afferents by progressively desensitizing the neurone each time one of its inputs fires (Thorpe & Gautrais, 1998). This could be achieved by a relatively simple mechanism involving feed-forward shunting inhibition in which all afferents reduce the sensitivity of the target cell, irrespective of their synaptic weights.

More precisely, let $A = \{ a_1 , a_2 , a_3 \ldots a_{m-1} , a_m \}$ be the ensemble of afferent neurones of neurone i, with $W = \{ w_{1,i} , w_{2,i} , w_{3,i} \ldots w_{m-1,i} , w_{m,i} \}$ the weights of the m corresponding connections; let $\mathrm{mod} \in]0,1[$ be an arbitrary modulation factor. The activation level of neurone i at time t is given by

$$\mathrm{Activation}(i,t) = \sum_{j \in [1,m]} \mathrm{mod}^{\mathrm{order}(a_j)} w_{j,i}$$

where order(a_j) is the firing rank of neurone a_j in the ensemble A. By convention, order(a_j) = +∞ if neurone a_j has not fired at time t, setting the corresponding term in the above sum to zero.

Neurone i will fire at time t if (and only if)

$$\mathrm{Activation}(i,t) \geq \mathrm{Threshold}(i)$$

Under such conditions, two important features can be pointed out :

• the better the match between the order in which afferent spikes arrive and the pattern of connectivity, the more strongly the neurone will be activated. Specifically, optimal activation is achieved when the spikes arrive in the order of their weights, with the inputs having the highest weights arriving first.

• the most strongly activated neurones (i.e. those where the order of their inputs best matches their weights) will tend to reach threshold and fire earlier.

These points are of great importance for further computation and learning.

2.2. The face detection model.

The most important feature in designing our model was to keep it as simple as possible, so that each step of processing could be fully understood. At the same time we wanted the architecture to be at least inspired by the first stages of processing in the visual system.

Thus, the architecture that we used was a very simple four-layer feed-forward neural network (fig. 2). Each layer was composed of a set of maps of different selectivity. Each map contained many neurones, each coding the specific information relative to a particular location (pixel) in the input image.

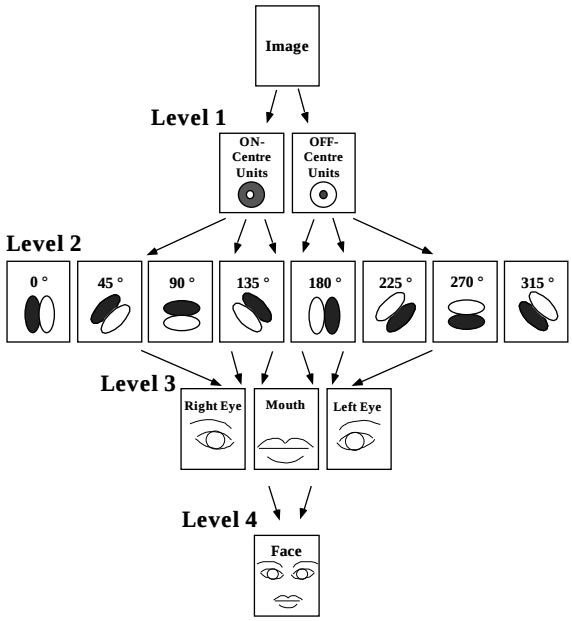

Figure 2. Architecture of the model. Connections are feed-forward only. The original image is first decomposed in two (positive and negative) local contrast maps, then in 8 different orientation maps (each separated by 45°). Units of the third layer respond to the activity in layer 2 that is characteristic of a mouth, right or left eye. Units at level 4 respond optimally when these 3 features are present simultaneously in the appropriate locations.

As in the primate visual system, receptive fields properties became increasingly complex as processing progresses. Units in the first layer had concentric ON- and OFF-centre receptive fields, like retinal ganglion cells, responding optimally to a positive or negative local contrast, whereas units in the second layer had orientation selectivity (8 different orientations, separate from each other by 45 °) similar to that seen in simple cells in V1. Units in the third layer were trained to respond to the pattern of activation in layer 2 that was characteristic of basic facial features (left eye, right eye and mouth) of a particular size, whereas fourth layer units were designed to respond optimally when these three different components were present in the appropriate locations, i.e. when a face was present. Thus the latter units had a pattern selectivity similar to that found in some neurones of the inferotemporal cortex (Bruce et al., 1981; Perrett et al., 1982; Abbott et al., 1996).

3. Learning Method and Receptive Fields.

3.1. Learning Method.

The principle of the learning method is based on the intrinsic properties of integrate-and-fire neurones and Rank Order Coding. As shown in fig. 1, the relative order of firing in cells of a population constitutes the code of the input stimulus. A neurone will reach its highest activation level when the relative order of firing in its inputs will best match the order of the corresponding connections (see section 2.1.).

Therefore, to detect a specific firing pattern in a population of simulated integrate-and-fire neurones, it is sufficient to use a set of connections that respect the relative order of this firing pattern. For instance, that order could be obtained by computing the mean order of a set of training examples.

3.2. Receptive fields.

3.2.1. Local contrast maps.

The first processing step in our model is the decomposition of the input image in local positive or negative contrasts. This can be achieved very easily using sharp gaussian laplacian filters. However, whereas in a more conventional neural network architecture, the value resulting from this convolution would be sent to later stages, in

our model, the resulting value is used to determine at what latency each cell will fire. The earliest firing cells in level 1 will thus correspond to the parts of the input image where the local contrast is highest.

3.2.2. Orientation maps.

At the second level of the computation, the input image is segmented in eight different orientation maps, each selective to a particular orientation. The receptive fields used for that purpose (fig. 3) are oriented Gabor filters of the same type as those used by Thorpe & Gautrais (1997).

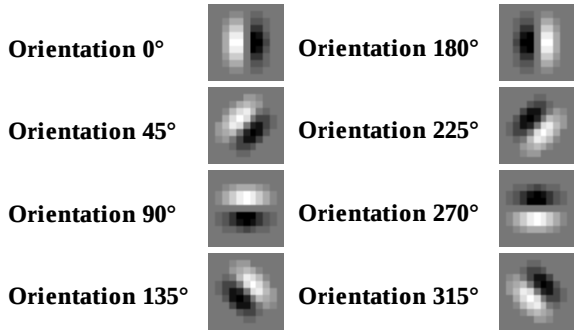

Figure 3. Receptive field organisation for orientation selective cells in layer 2. Bright and dark pixels correspond respectively to positive and negative weights from ON-centre cells in layer 1. Connections from the OFF-centre cells map are identical, but rotated by 180°.

Neurones at that level show selectivity in that they respond at shorter latencies when the orientation of an edge in the image matches the shape of their receptive fields.

3.2.3. Feature selective maps.

The patterns of connections between the eight orientation-tuned maps in layer 2 and the feature detecting cells in layer 3 were set in such a way that the cells responded best when the order of activation in the different maps was close to that seen with a set of training stimuli.

We used a training database of 270 front (± 30 °) views (92x112 grey-level images) of male and female faces (10 views of 27 persons), of which only a small proportion wore glasses (2%) or had a beard (11%). For each image, the precise locations of the right and left eyes and mouth were determined manually. The images were propagated through layers 1 and 2 of the network, and a region of the appropriate size, around the location of the mouth or left or right eyes, was extracted from each orientation map. The size of this region was determined so that it should include not only the feature itself, but also the immediately surrounding area. Thus the receptive fields of the "eye-detecting" neurones included the eyebrows whereas the zone of interest for the "mouth-detecting" cells included part of the nose. For each feature (mouth, left and right eyes) and for each orientation map, the mean order of firing in the corresponding region was computed over the entire database.

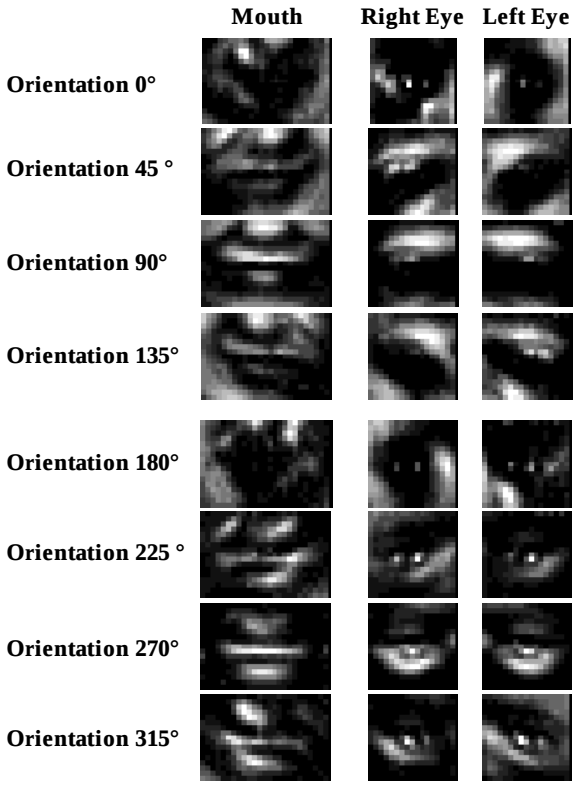

Figure 4. Receptive field organisation of feature selective neurones. All weights are positive, with black pixels set to zero and white ones corresponding to a maximum weight value. Each orientation map is connected to each feature selective map by the corresponding weight set.

The resulting mean order patterns were then used directly to determine the strength of the connections linking the orientation and feature maps (fig. 4). As a result, neurones in each feature-detection will be strongly activated only if the corresponding

feature is present at the appropriate location. Thus the position of the firing neurones in such a map gives information about the precise location(s) of the feature(s) in the input image.

3.2.4. Face detection map.

Neurones in the level 4 (face-detection map) were set up to fire if the three basic facial features (mouth, left and right eyes) are present in the image, and in the appropriate locations. This "facial structure", i.e. the relative positions of the component features of a face, can be easily described using a set of three gaussian filters centred at the appropriate positions in the receptive fields of such a neurone (fig. 5).

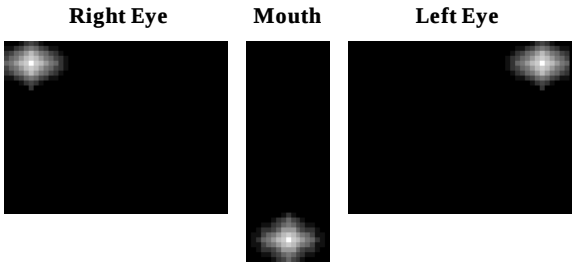

Figure 5. Connection patterns between feature detecting cells in layer 3 and the face selective neurones in layer 4. Each of these filters is centred at the neurone's location.

4. Results.

4.1. Propagation results.

The pattern of firing obtained when an image containing a face is propagated through the network is illustrated in figure 6. For clarity, only 4 of the 8 orientation tuned maps are shown. Within each map, the brightness of the individual points corresponds to the order in which the neurones fired - bright spots correspond to neurones that were among the first to fire and the grey scale value gets progressively darker for later firing neurones.

It is clear that the network performs accurately. Activity in the three feature level maps is restricted to the places corresponding to the locations of the right eye, the mouth and the left eye respectively. Similarly the region of activated cells in the face level map corresponds to the centre of the face.

It should be noted that such a network has no problem coping with an input image containing more than one face. The large number of neurones in the face-detection layer means that there is effectively one "face-cell" for every pixel in the image. Images with multiple targets will simply produce the appropriate number of activated regions in the output map.

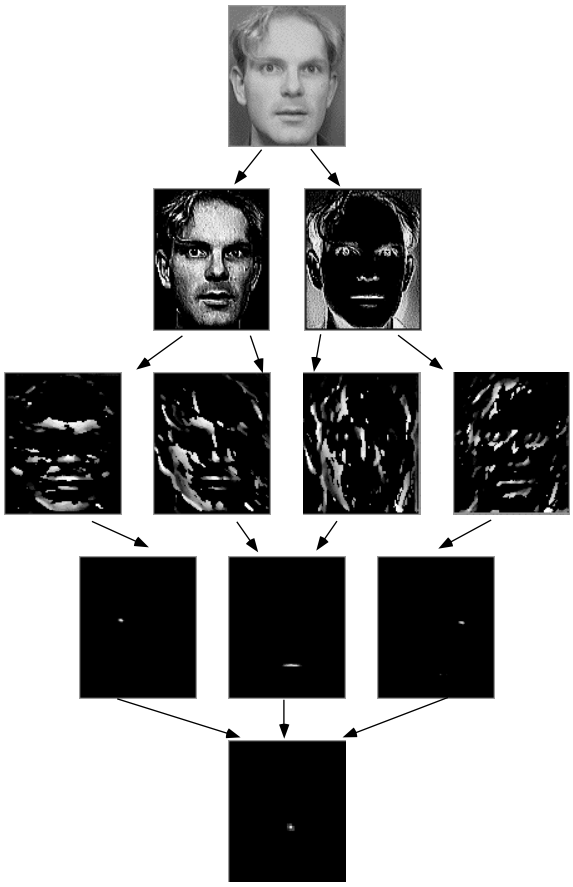

Figure 6. Result maps obtained after the propagation of an image of a face in the network. From top to bottom : the original image, the ON- and OFF-centre cells maps, 4 out of 8 orientation maps, the 3 feature-detection maps (right eye, mouth and left eye) and the face detection map. In each map, non-black pixels correspond to firing neurones, with the grey-level intensity representing the firing order in the entire layer (neurones that fired first are brighter). The position of the firing neurones in the last layer gives the explicit location of the face in the original image.

Conversely, when an image containing no face is propagated through the network,

the first two computation steps are run as described above, leading to a representation of the image in terms of oriented edges. However, none of the feature selective neurones in the following layer should receive an input ordered well enough to let it reach its threshold. As a result, no facial features should be detected and no face-cell activity will be present in the final layer.

4.2. Statistical results.

4.2.1. Testing method.

To evaluate the performance of the model, we have tested it with a large range of natural images. Two public databases containing many face images were used, together with our own database of natural images that had no face present.

The first database was obtained from the Olivetti Research Laboratory and consisted of 400 frontal (± 30 °) views (92x112 grey-level images) of male and female faces (10 views of 40 persons), that we separated in 2 groups. The first group, which we will call database 1, included the 270 images that were used for training and contained only a small percentage of people wearing glasses or with a beard. The second group ("database 2") contained the remaining 130 images (10 views of 13 persons), and had a large proportion of people wearing glasses (88%) or a beard (31%). These were used as a set of "difficult" examples.

The next database, which we will call database 3, contained 300 frontal views (256x171 grey-level images) of different people whose faces were approximately the same size as those used for training. Roughly half of them wore glasses, while 16.6 % had a beard, proportions that can be considered as reflecting the "every-day life" conditions.

The last database ("database 4"), contained 216 (84x104) grey-level images of natural scenes with no faces, but a large range of animals, plants, landscapes or objects such as cars, buildings and food... It was used to determine the error rate of the model.

We define the detection rate as

$$\text{Detection rate} = \frac{\text{Number of detected features/faces}}{\text{Number of features/faces present in the database}}$$

and the false detection number as

$$\text{False detection number} = \text{Number of firing regions where the feature/face was not present}$$

Half of the images of the learning base (database 1) were used to determine optimal values for the various parameters of the system, which include the modulation value and the threshold levels. These were set individually for each feature map in order to maximise the detection rate whilst keeping the mean false detection number at below a maximum of 1 false detection per image. Coefficients for the face-selective neurones were chosen so that the combination of at least two of the three features should be present in the correct locations to make the neurone fire.

For each of the other databases, each image was propagated through the network. The detection rate and the false detection number were then determined for each detection map.

4.2.2. Testing results.

Detection Map → Test Database ↓	**Mouth**	**Right Eye**	**Left Eye**	**Face**
Database 1 (135 images)	92.3 % 18	97.8% 97	95.6% 100	**96.3% 2**
Database 2 (130 images)	88.5 % 27	83.1% 88	80% 87	**73.1% 4**
Database 3 (300 images)	91 % 89	92.7% 222	75% 198	**94 % 4**
Database 4 (216 images)	- 14	- 13	- 12	**- 1**

Table 1. Results of the model with four different test databases. For each database and each detection map, the top number indicates the detection rate, the bottom number indicates the false detection number. This number must be compared with the number of images in the database, and the number of neurones per image. Database 2 is the « difficult » example base. Database 4 contains no face image.

The results obtained are shown in table 1. Results with database 1 indicate that the

faces were accurately located more than 95% of the time, with very few false detections (an estimation of the rate of wrongly firing neurones would be lower than 0.001%). The detection rates of the feature-selective neurones are also very high, whereas the number of false detection is still low, when compared with the number of images (135 in this database) and the number of feature-selective neurones per image (10304 for each feature).

Results with database 2 can be considered as minimal detection results since the faces were particularly difficult to detect : 88% of the people in the database wore glasses, which probably explains the decrease in the detection rate for the eyes, while 31% had a beard, which probably led to a decrease in the mouth detection results. Thus the 73.1% detection rate can probably be considered as the model's minimal performance.

This is corroborated by the results obtained with database 3, which demonstrate a surprisingly good ability to generalise to novel faces, since the face detection rate is approximately the same as that obtained with the learning base. It is worth noting that the image contrast and illumination conditions were very different in these 2 databases, as can be seen from the marked decrease in the left eye detection rate with database 3. Nevertheless, such differences in image quality do not seem to disrupt the performance of our face detection model. Furthermore, the detection rate was not much influenced by the substantial numbers of people in the database that wore glasses (50 %) or had a beard (16.6 %).

Finally, results on database 4 clearly demonstrate the specificity of the network responses : when no face is present in an image, the probability of a false feature detection at any particular location is very low (about 0.002%), while the probability of a false face detection is virtually equal to zero.

Thus the specificity of the simulated face-selective neurones responses can be summarised as follow :

- when a face is present in such a neurone's receptive field, the probability of a response is about 95%.
- when no face is present in this receptive field, the probability of a response is virtually zero.

5. Discussion.

Using a very simple four-layer feed-forward neural network model, we have been able to show that only one spike per neurone is sufficient to perform quite complex processing tasks such as face detection and localisation in a natural images.

Furthermore, the results obtained appear to be remarkably good by comparison with more classic image processing techniques. Principal components analysis (Turk & Pentland, 1991), which seems to be one of the most widely used methods for face detection (Valentin et al., 1994), does not lead to significantly better results than ours. For example, Sung & Poggio (1994) and Moghaddam & Pentland (1995) both achieved a 90 % detection rate, with a false detection rate that was not significantly better than the one obtained here.

From a purely practical point of view, modelling visual processing with SpikeNET has advantages over more classical methods that stem from its computational efficiency. One of its main features is that it is "event-driven" - the main task of the simulator is simply to propagate spikes. As a result, if there are no spikes in a particular layer, then there is no computation to do. This is particularly important in multilayer feed-forward networks where higher levels in the hierarchy are not involved at all until relatively late in the propagation.

In addition, once a spike has been emitted in our model's first layer, it is immediately processed by the following layer, whose neurones can fire spikes in their turn, and so on. A given layer does not need to wait until all the preceding layer's neurones activities have been calculated before it can start computing. That is a very original feature of the asynchronous propagation used here and which distinguishes it radically from classical neural networks.

Furthermore, the thresholding mechanism provides another way of minimising computation time. With relatively low thresholds, a neurone can fire when only a relatively small fraction of its inputs have fired (on condition, of course, that the ones that do fire early have high weights). Thus face-selective neurones in the last processing layer can sometimes emit a spike when only 20% or less of the cells in layer 1 have fired. Of course, this will only be possible when the face in the image is particularly clear.

Since the first spikes occurring in any given layer correspond to the neurones that were first to reach their threshold, the asynchronous mode of transmission used here guarantees that the most salient information is computed first.

All these dynamic features, which are based on the observation of information propagation in the visual system, mean that our model tend to be much faster than classical neural networks.

For instance Rowley et al. (1995) describe a neural network model for localising faces in an image, with a detection rate of roughly 90% and a false detection rate of about 0.0002%, that took roughly 6 minutes on a Sun SparcStation 20. Running on a simple PowerPC 604e at 150 MHz, the network presented here was able to localise the faces in an image with a processing time ranging from 1 to 5 seconds (depending on the number of firing neurones and the size of the image).

The simple four-level architecture used here is certainly a very poor description of the real visual system, and we certainly would not wish to claim that the perception of faces in humans and monkeys can be realistically modelled in such a primitive way. It is clear that the primate visual system involves extensive feed-back connections at virtually every stage as well as a large number of horizontal connections. We are currently exploring the situations in which such connections could play a role.

One major simplification of the present model is that we assume that both the activation levels and the sensitivities of all the neurones in the network are reset before each stimulus is presented. Clearly, this is not a realistic assumption. Allowing the activation values of neurones in the network to start from random values would certainly increase noise in the rank order code. However, as we have shown elsewhere (Thorpe & Gautrais, 1998), the rank-order coding scheme is remarkably selective and the probability of a neurone responding to a random sequence of inputs can be made very low. Furthermore, it may be that under natural viewing conditions, the suppression of thalamic transmission that occurs during every saccade may mean that there is indeed a form of reset every time a saccade occurs.

Nevertheless, the main claim that we would wish to make on the basis of these simulations is that visual processing based on one spike per neurone is indeed a possibility. In real visual systems, it is obvious that neurones do, in fact, typically generate trains of spikes in response to a given stimulus. As a result, it would be difficult if not impossible to exclude a role for rate coded information. However, in the sort of simulated visual system that can be explored using SpikeNET, restricting each neurone to one spike and one spike only appears perfectly feasible. Under those conditions it becomes possible to demonstrate that rate coding is not required for visual processing.

References.

Abbott, L. F., Rolls, E. T. & Tovee, M. J., 1996, Representational capacity of face coding in monkeys. Cerebral Cortex, 6, 498-505.

Bruce, C. J., Desimone, R. and Gross, C. G., 1981, Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology, 46, 369-384.

Gautrais, J. and Thorpe, S. J., 1998, Rate coding vs. temporal order coding : a theoretical approach. Biosystems, XXX, XXX-XXX.

Jeffreys, D. A., 1996, Evoked potential studies of face and object processing. Visual Cognition, 3, 1-38.

Moghaddam, B. and Pentland, A., 1995, Probabilistic visual learning for object detection. In The Fifth International Conference on Computer Vision. Cambridge, MA.

Nowak, L. G. and Bullier, J., 1997, The timing of information transfer in the visual system. In J. Kaas, K. Rocklund, and A. Peters, Eds.), Extrastriate cortex in primates, pp. sous presse, Plenum Press.

Oram M. W. and Perrett D. I., 1992, Time Course of Neural Responses Discriminating Different Views of the Face and Head. J Neurophysiol, 68, 70-84.

Perrett, D. I., Rolls, E. T. and Caan, W., 1982, Visual neurons responsive to faces in the monkey temporal cortex. Experimental Brain Research, 47, 329-342.

Rowley, H. A., Baluja, S. and Kanade, T., 1995, Human face detection in visual scenes., Internal Report, School of Computer Science) Carnegie Mellon University, Pittsburg.

Sung, K. and Poggio, T., 1994, Example-based learning for view-based human face detection. Proceedings Image Understanding Workshop, II, 843-850.

Thorpe, S. J., 1990, Spike arrival times: A highly efficient coding scheme for neural networks, in: Parallel processing in neural systems, R. Eckmiller, G. Hartman and G. Hauske (eds.), pp. 91-94, North-Holland: Elsevier.

Thorpe, S. J. and Gautrais, J., 1997, Rapid visual processing using spike asynchrony, in: Neural Information Processing Systems 9, M. C. Mozer, M.I. Jordan and T. Petsche (eds.) (MIT Press, Cambridge) pp. 901-907.

Thorpe, S. J. and Gautrais, J., 1998, Rank Order Coding : A new coding scheme for rapid processing in neural networks, in: Computational Neuroscience : Trends in Research, J. Bower, (ed.) (Plenum Press : New York).

Turk, M. and Pentland, A., 1991, Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3, 71-86.

Valentin, D., Abdi, H., O'Toole, A. and Cottrell, G. W., 1994, Connexionnist models of face processing : a survey. Pattern Recognition, 27, 1209-1230.