Classification of Facial Expressions with Domain Gaussian RBF Networks

James M. Hogan, Michael Norris and Joachim Diederich*
Machine Learning Research Centre
Faculty of Information Technology
Queensland University of Technology
{hogan,michaeln,joachim}@fit.qut.edu.au

Abstract.
This paper examines the problem of categorisation of facial expressions through the use of a receptive field neural network model, based upon novel domain Gaussian network units trained through error backpropagation. Such networks are trained upon images derived from the Ekman and Friesen “Pictures of Facial Affect” database, and they are subsequently able to successfully generalise to images of unseen subjects, and provide qualitative replication of the perceptual confusions common to previous studies. By using digital morphing techniques to produce intermediate frames between the existing stills, we are able to study the space of transitions between endpoint expressions, our results suggest that expressions unrelated to the endpoint images may be perceived during certain transitions, a path far more complex than direct translation through a neutral expression.

1. Introduction

This work is concerned with the application of novel neural network models to the problem of categorising human facial expressions, and the extent to which such models may successfully replicate human performance and provide falsifiable predictions for novel images. The human category judgments are here obtained from the now classical Ekman and Friesen “Pictures of Facial Affect” study [3], the still images of their set of photographs partitioned by Ekman’s subjects into one of the six classes of “Happiness”, “Sadness”, “Anger”, “Fear”, “Surprise” and “Disgust”. The task of each of our networks is to recognise one of these categories from training upon positive and negative examples of that category, these consisting of pre-processed bitmaps derived from scanned versions of the Ekman stills.

As well as using Ekman images unseen by the network for generalisation testing, we also examine network performance on novel images extracted as still pictures from a number of (linear) digital morphs between images derived from the Ekman slides. In this way we are able to identify both the crossover point in the transition ‘movie’ – the frame at which the final category activation exceeds that for the initial category – and to offer psychologically interesting predictions of perceptual confusion, with some sequences appearing to verge during the transition toward a unrelated category before returning to the appropriate endpoint.

*This work was funded by a grant from the Australian Research Council.
The network is based upon novel receptive field units which map facial regions implicated in the encoding of expressions, allowing generalisation to unseen subjects and some confidence measure of the degree of category membership for each image. These domain Gaussian units provide a strong weighting of the intensity of two-dimensional features within the image according to their distance from the nominal centre of the field, giving a sensitive but compact representation of local changes within the domain.

From the cognitive perspective, our model is closely allied with the Lzero project of Feldman et. al. [4] in addressing the integration of linguistic and visual information. In the original Lzero task, a computational system is required to learn a small portion of a natural language – as defined by a set of associations between spatial relations terms such as above and cartoon pictures exemplifying each term – sufficiently well that it is able to judge whether an unfamiliar picture is an exemplar of a particular term. In the present work, we are able to establish an association between the semantics of terms such as “happy”, and their denotation through the relevant facial expression, which we argue is identifiable through a specific spatial configuration within the image.

This paper is organised as follows. Section 2 provides a brief summary of the problem of facial expression perception, with particular reference to the the relevant developmental literature. Section 3 discusses the network model and the receptive field units employed for each network, together with an account of the adaptive learning approach of Geva [5], which forms the basis of the training methods employed in this study. Details of the image preparation and the precise learning task are given in section 4, and are followed by presentation of the learning and generalisation results for unseen static images in section 5. Section 6 examines the creation of transition movies between well-classified Ekman stills, and provides an initial characterisation of the (potentially) complex space of transitions between categories. After discussion of these results in section 7, we summarise our conclusions in section 8.

2. Development of Facial Expression Perception

Perception of the human face has been of great scientific interest since the time of Galton, and categorisation remains a central question in this research – notably in regard to the recognition of a familiar subject among distractors. In the present context, it is useful to consider both the development of sophisticated facial processing and the parallel development of language processing.

Young infants are known to scan faces on the basis of individual salient features [9], [6], [16], showing a marked preference for tracking facial stimuli as early as the first hour of life, suggesting some innate preference for conspecifics, which is refined to the infant’s caregiver once visual acuity has developed sufficiently. This preference continues throughout the first month of life, subsequently declining significantly, only to re-emerge around a month later as a result of the maturation of primary visual cortex [9]. Corresponding development of the fovea and the resulting acuity allow greater resolution of facial components during this period, and by five months, a preference for facial stimuli is shown only if movement of constituent features is integrated into the pattern display [9]. One example of this is a minor adjustment to a line-drawn mouth. Hainline [6], reports further development of scanning from 7 weeks through to the establishment of an integrated facial percept by the end of the first year of life.

Studies of changes in performance on facial expression tasks with increasing age are limited, and variations in experimental approach make it difficult to trace development of this capability right through to adulthood. Some indication may be gleaned from the
related case of face recognition, in which there is evidence of a near-monotonic improvement in performance from around five years to maturity [1]. What is clear, however, is that by age 4-6, children may be trained to associate verbal descriptions (such as “happy”) with line drawings of appropriate facial expressions [12], and that this capacity develops over time, being fully established well before adolescence (Dimitrovsky et. al. [2]). However, this latter study also reports significant variations in the confusion rate across the categories, and in performance with respect to age among the more difficult judgments. Chung and Thomson [1] postulate that improved recognition performance may be the result of memorisation of a greater number of facial features and feature relations, rather than involving a fundamental change in mechanism.

3. Radial Basis Functions

Networks of radial basis function units [13] have been applied to a wide variety of classification and function approximation tasks. While in theory the approach provides no representational advantage over sigmoid networks\(^1\), their strongly localised response offers some simplification of network construction and training, as learning need not depend upon a fine balance between units strongly active across much of the domain [5].

The RBF unit response to a particular input varies according to the (Euclidean) separation between the input vector and the unit centre, both being regarded as elements of the vector space of patterns. Such responses are usually weighted according to the spatial Gaussian of equation 1, with influence decaying as the separation increases, but remaining strong within some local radius of the centre. Formally, given an input vector \( x \in \mathbb{R}^n \), and a centre \( c \in \mathbb{R}^n \), a Gaussian RBF unit in the general form given by Geva [5] has activation

\[
G(x; T, c) = \exp(-||T(x - c)||_2^2),
\]

where \( T \in \mathbb{R}^{n\times n} \) is the transformation matrix, which determines the shape of the unit’s footprint. Usually \( T \) is chosen to be diagonal

\[
T = \text{diag}(1/r_1, \ldots, 1/r_n),
\]

\( r_i \) being the local radius in direction \( e_i \), the \( i \)-th component of the standard Euclidean basis. In the simplest case, \( r_i = r, \forall i \), producing a radially symmetric receptive field.

3.1. Domain Response Units

The use of RBF units such as those described above for problems involving 2D images leads to a training problem of enormous dimension. For this reason, some mechanism for incorporating the influence of an entire region – without \( \text{a priori} \) geometric specification – must be found if bitmap representations are to be employed. Gaussian Domain Response units, introduced below, are one such approach.

Formally, we consider an image over a bounded domain \( D \subset \mathbb{R}^2 \) with normalised intensity level:\(^2\)

\[
I(x) : D \to [0, 1],
\]

\(^1\)Hornik et. al. [8] established that a two layer feedforward network of sigmoid neurons may approximate any continuous function \( f : \mathbb{R}^n \to \mathbb{R}^m \), and a similar result for RBF units was established more recently by Hartmann et. al. [7].

\(^2\)This need not be restricted to intensity and may include any suitably normalised scalar combination of image components.
at each point $x \in D$ within the region. Then the activation of the Domain Response Unit centred at $c$, with transformation matrix $T$, is defined by:

$$v(I; T, c) = \frac{1}{<G>} \int_D I(x) G(x; T, c) dx,$$  \hspace{1cm} (4)

where

$$<G> = \int_D G(x; T, c) dx,$$  \hspace{1cm} (5)

is the maximum unit response.

In this work we are concerned only with a 2D image map of dimension $N^2$, each location having identical nominal area, with constant intensity $I_{ij} = I(x_i, y_j)$ across this logical pixel. Equation 4 for the domain response is then re-cast to give

$$v(I; T, c) = \frac{1}{<G>} \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij} G(x_i, y_j; T, c),$$  \hspace{1cm} (6)

with the corresponding modification to $<G>$. Here $(x_i, y_j)$ is understood to be the centre of the logical pixel $(i, j)$.

The network is trained according to the parameter-specific dynamic learning rate approach described in the next section, unit centres and radii being adjusted through backpropagation of error to provide receptive fields appropriate to the problem. While calculation of the activation for each unit is time-consuming, network training is less troublesome because of the dramatic reduction in the dimension of the search space.

3.2. Network Training Methods

Geva [5] investigated the use of fully connected RBF networks for function approximation, a layer of RBF units receiving activation from the inputs, and feeding a weighted response to a linear output unit to produce the function estimate. In contrast to the more usual clustering techniques\(^3\), Geva’s approach provides receptive fields appropriate to the problem through error backpropagation adjustment of the unit centres and radii, sharing with the traditional methods gradient descent adjustment of the output weights.

Geva’s work significantly extended the existing approaches in two key areas: through the generalisation of the response function described above, and through the development of a training technique based upon the adaptive gradient descent method of Silva and Almeida [15]. Traditional neural network function approximation techniques – based upon backpropagation-trained multi-layer perceptrons – have long suffered from convergence problems in larger domains. One contributing factor to this problem is the use of a single global learning rate in spite of large variations in the magnitude of partial derivatives with respect to particular weights. The method of Silva and Almeida addresses this issue by providing for the adjustment of each weight through a learning rate set with reference to the relevant partial derivative.

While Geva remarks that Silva and Almeida’s technique is unable to overcome the difficulties posed by a number of large MLP problems, he nevertheless establishes that the technique is well-suited to the training of an exponential response network, perhaps because of the purely local effect of parameter adjustment [5].

As is commonly the case in backpropagation approaches, the estimation error is given by the squared error cost function of equation 8, with contributions summed over

\(^3\)See for example [11].
the entire set of $P$ patterns. For each pattern, the true activation value on pattern $p$, $y_p$, is approximated by the network estimate

$$
\hat{y}_p = \sum_{\mu=1}^{m} w_{\mu} v_{\mu},
$$

(7)

where $v_{\mu}$ is the output of domain response unit $\mu$ on pattern $p$, and $w_{\mu}$ the corresponding output layer weight.

Reliance on individual learning rates requires that we provide expressions for the partial derivatives of the cost function with respect to each of the trainable parameters. The necessary expressions for a two dimensional map are developed as follows. As discussed above, the squared error cost function for the problem is defined by the equation

$$
E = \frac{1}{4} \sum_{p=1}^{P} (\hat{y}_p - y_p)^2,
$$

(8)

where the constant is chosen to provide cleaner derivative expressions. The derivatives with respect to output weights are then

$$
\frac{\partial E}{\partial w_{\mu}} = \frac{1}{2} \sum_{p=1}^{P} (\hat{y}_p - y_p)^2 v_{\mu}(I_p; c^\mu, T^\mu),
$$

(9)

where as before $v_{\mu}$ is the activation of the $\mu$th domain response unit. Use of the chain rule and some manipulation yields the remaining expressions for the partials with respect to the transformation matrix elements

$$
\frac{\partial E}{\partial T_1^\mu} = \frac{-T_1^\mu}{<G>} \sum_{p=1}^{P} (\hat{y}_p - y_p) w_{\mu} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^p G(x_i, y_j; T^\mu, c^\mu)(x_i - c_1^\mu)^2 \right),
$$

(10)

and

$$
\frac{\partial E}{\partial T_2^\mu} = \frac{-T_2^\mu}{<G>} \sum_{p=1}^{P} (\hat{y}_p - y_p) w_{\mu} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^p G(x_i, y_j; T^\mu, c^\mu)(y_j - c_2^\mu)^2 \right);
$$

(11)

together with those with respect to the field centres

$$
\frac{\partial E}{\partial c_1^\mu} = \frac{(T_1^\mu)^2}{<G>} \sum_{p=1}^{P} (\hat{y}_p - y_p) w_{\mu} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^p G(x_i, y_j; T^\mu, c^\mu)(x_i - c_1^\mu) \right),
$$

(12)

and

$$
\frac{\partial E}{\partial c_2^\mu} = \frac{(T_2^\mu)^2}{<G>} \sum_{p=1}^{P} (\hat{y}_p - y_p) w_{\mu} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} I_{ij}^p G(x_i, y_j; T^\mu, c^\mu)(y_j - c_2^\mu) \right).
$$

(13)

As one would expect, these expressions are substantially more complex than those for Geva’s units of equation 1, the difference arising as a result of the summation over the pixels required to calculate the unit activation. However, domain response units share with more conventional RBF units the problem of large variations in gradient magnitudes – both through local variations in the pattern space and through substantial differences in functional form between the field centre and radius partials. Fortunately, the individual learning rates $\eta$ may be successfully tuned for each of the adjustable weights, with modifications to the strategy described by Geva being limited to the initial unit array (see section 4).
3.3. *A Biological Interpretation*

While the properties of the domain Gaussian representation are of themselves desirable, we may also view the units as a computationally convenient form of a biologically plausible receptive field, arising as a result of variations in the density of synaptic contacts. It has long been known \[14\], that the probability of synaptic connection is high in the immediate vicinity of each neuron, but is observed to decrease markedly with increasing spatial separation between the ‘source’ and ‘target’ cells. Let us now consider a target neuron, existing at the centre, \(c\), of a logical ‘canopy’ over a large but finite group of randomly positioned source nodes as shown in figure 1, such as might be the case if the target node resided in a different layer from the source nodes. Suppose that each of the source nodes projects a connection of fixed (and perhaps unit) weight to the target with probability \(\alpha(r)\), where \(r\) is the radial separation between the source node at \(x\) and the target at \(c\) as normalised with respect to some local radius \(R\), and \(\alpha\) is a non-increasing function of \(r^4\).

If we define a connectivity function \(C(r)\) to be the cumulative fraction of connections to the target within radius \(r\) of the centre, the fraction contained within the area element of figure 2 may be expressed through the equation

\[
\{C(r + dr) - C(r)\}d\theta = r\alpha(r)drd\theta. \tag{14}
\]

The corresponding connection density \(c(r)\) is then

\[
c(r) = \frac{dC}{dr} = r\alpha(r), \tag{15}
\]

where \(\alpha\) must be chosen so that

\[
\int_0^\infty c(r)dr = 1. \tag{16}
\]

If we make the mathematically convenient selection of

\[
\alpha(r) = e^{-r^2}, \tag{17}
\]

\(^4\text{There may of course be substantial angular variation in connection probability – corresponding in some respects to the elliptical footprint of the weighting function of equation 1. For the sake of simplicity, we shall restrict our discussion to radially symmetric connection probability \(\alpha(r)\) and intensity \(I(r)\).}\)
then the connection density is seen to be a weighting function for the map, with the effective influence of each feature upon the receptive field critically dependent upon its separation from the unit centre. The field response is then formally the expected value of the intensity, weighted with respect to the connection density across the domain

\[ v(c) = \int_0^\infty I(r)c(r)dr, \]  

(18)

with properties qualitatively similar to the more general response of equation 4.

4. The Learning Task

In the present study, the 110 images of the “Pictures of Facial Affect” database were scanned from photographic slides to gray-scale pixel images and preprocessed to provide some alignment of facial features between frames, and to eliminate confounding detail. In all cases, the preprocessing operations have an identifiable counterpart within the visual system, and our aim was limited to that of reducing the scope of the facial expression task to that of comparison of images with facial features in gross-level correspondence. Each pixel map was manually scaled (preserving proportions) and cropped so that the face from chin to hairline and ear to ear fit as closely as possible within a 3 by 4 rectangle. The resulting images were bandpass filtered (using Photoshop 4.0)
under a Gaussian blur with radius 3 pixels, followed by a high pass filter with radius 10 pixels. This filtering serves to remove fine textural detail and normalise across complexion, lighting, and hair colour. Images were then reduced using Gaussian compression to 30 by 30 pixels, a scale at which the facial expressions remain distinguishable by eye.

The resulting images were presented as input to the neural network models. Targets corresponding to each non-neutral expression were used for training, with target values of 1.0 indicating category membership (as judged by Ekman’s human subjects), and 0.0 indicating non-membership. Each model network was initialised with 64 equally-spaced receptive fields which under training formed a small set of highly localised and redundant receptive regions.

For each of the six target expressions – “Happiness”, “Sadness,” “Fear”, “Anger”, “Surprise” and “Disgust” – eight separate networks were trained, with the best performed network in each category preserved as the ‘judge’ network. Such networks are subsequently referred to by their category name, respectively the “Happy” network, the “Sad” network and so on. An overall category judgment may be made through a winner-take-all across the networks.

Due to the paucity of data, the nets are trained on a leave-one-subject out approach, in which all of the six images of a single Ekman subject are reserved for testing, each network being trained on the the remaining 104 frames.

5. Results for Static Images

Examples of raw generalisation results are shown for unseen subjects NR, JJ, EM and GS, MF and MO, and JM in tables 1 through 5. Each of the judge networks was trained on all the images remaining in the database, for 500 epochs in the case of NR and JJ, and 300 epochs for the remainder. As may be seen from the activations, the present task provides substantial variation in prototypicality across subjects, with some individual faces being naturally biased – or perhaps biased as a result of the limited technical proficiency of the actor – towards some categories and away from others. For example, the Happy pose (#14) for subject EM produces an activation of 0.96 in the relevant Happy network, while the corresponding task for GS (#22), yields only 0.55. Thus, absolute thresholds of classification performance are an unreasonably demanding measure in the present context, and we consider that the judge networks have generalised successfully to the unseen subject if the highest activation for a particular category network (for example the Anger network for unseen JJ) occurs on the unseen image of that category (here the Angry JJ or image 38). Note that while the example may show correct generalisation, this does not of itself imply that the response of the appropriate category network (for #38: Anger: 0.50) must be the highest for that image (for #38: Disgust: 0.59). Indeed such confusions are part and parcel of the problem, with the Ekman and Friesen study reporting confusion of a significant extent for a number of examples of Fear and Surprise, of Anger and Disgust, and to a lesser degree between Anger and Sadness.

While the networks of the present study do not thoroughly replicate the performance of Ekman’s judges, they nevertheless display similar behaviour, with a more pronounced confusion in the ‘perception’ of Sadness. More detailed discussion of the performance on example networks is provided in the captions of tables 1 to 5.
6. Digital Morphing and Dynamic Images

The Ekman and Friesen study is concerned solely with the classification of single still frames. While these images may exhibit substantial variation within classes, as we have discussed in the previous section, they do not support direct investigation of the category boundaries, the point at which a perturbation of an image of one category is sufficient for that judgment to be revised to another. One way of examining this problem is to create a movie of each possible category transition for a number of subjects, and hence present the individual frames to the category networks for judgment.

For each of the example subjects discussed above, digital morphing software was used to produce a series of 16 intermediate frames between the well-classified Ekman stills. Key frames were high resolution scaled, cropped and filtered as described above, with the 34 morphing control points chosen manually. Linear interpolation of distorting control points and linear pixel value crossfade results in a natural and gradual transition between expressions. Due to the small number of control points used, some details – notably teeth and lips – were seen to merge unnaturally, although the integrity of the image as a face with a recognisable expression was not impaired.

Each of the resulting 15 animated sequences (a linear morph between each possible category pairing for a given Ekman subject) is then presented in parallel to all of the model networks. As one might expect, the networks representing the initial and final categories of the morph showed respectively decreasing and increasing activation levels. However, for some of the subjects, the responses of the remaining categories suggest a perceptually confusing trajectory may be followed as the transition takes place, with significant activation of a number of the judgment networks. This is in contrast to our intuitive expectation of a cartesian category space, with transitions constrained to pass linearly through a conceptual origin – although there is more than a hint of this phenomenon in the strong intermediate response of the Sadness network for subjects MF, JM and EM, with Sadness known to be confused with Neutral in some of the Ekman and Friesen experiments.

In the absence of an exhaustive study involving leave-one-subject-out networks and presentation of all possible sequences, we must thus be cautious in our remarks, noting only that our results suggest

- the existence of complex transitions in which one or more apparently unrelated categories may compete for apprehension during the intermediate phases;

- that there is consistent qualitative – and to some degree quantitative – grouping of the responses for Fear with those for Surprise, as well as those for Anger, Disgust and Sadness, although these may be disrupted for particularly strong exemplars of a particular category.

Figures 3 through 8 show the response of each of the relevant category networks to the image transition indicated, with a more detailed discussion of the particular results being given in the associated caption. The implications of these results, and those for the static case are considered in the following section.

7. Discussion

In this work we have considered the problem of classification of facial expressions for an unseen subject within the categories employed by Ekman and Friesen [3]. Both the
generalisation results of section 5 and the morphing studies of section 6 provide some evidence that these categories may be broadly distinguished through a decomposition of the form

\[ H_p \lor \{ F_r, S_r \} \lor \{ A_q, D_q, S_d \}. \]

This assertion is further supported by principal component analysis of the hidden unit activations, which suggests that as few as four dimensions may suffice to distinguish the categories. At the highest level, a decomposition into \( H_p \land \neg H_p \) is straightforward, with difficulty increasing as one attempts to distinguish between the Fear-Surprise and Anger-Disgust-Sadness sets, and finally between the elements themselves. In this latter case at least the targets are necessarily probabilistic, and while Ekman’s human judges report lower confusion rates than the domain Gaussian nets of this study, they are nevertheless significant within these distinctions.

Such indeterminacy is a fundamental aspect of human categorisation and the results of the present study are in some respects reminiscent of the confusion reported in Labov’s seminal ‘CUP’ study [10], in which subjects were required to attach linguistic labels to non-deterministically combined features of cups, mugs and bowls. More technically, Labov’s work was concerned with the fuzzy boundaries of denotational categories\(^5\), and the ‘CUP’ study demonstrated that small variations in the features underlying such denotata will contribute to a gestalt judgment of greater or lesser – or equivalently, more or less prototypical – membership of the category. At some point, such gradual variations give way to hard categorical distinctions, and even the use of controlled linear frame interpolation provides no guarantee that the resulting category judgments will be any more predictable.

8. Conclusions

In this paper we have introduced a new class of radial basis function unit and, through the exploitation of existing variable learning rate techniques, applied networks of such units to the problem of classification of facial expressions, as represented by the categories and images of the famous “Pictures of Facial Affect” study of Ekman and Friesen. Our models have successfully generalised to unseen subjects, providing clear classification within each subject while replicating the confusions of the earlier work. The novel use of morphing techniques to study the transitions between images suggests the existence of a complex perceptual transition space rather than a straightforward path through a conceptual origin. This space is the subject of ongoing computational and psychophysical investigation.

\(^5\)Denotation is defined as the association of a linguistic label with a set of possible referents, in the present work the association between an expression label such as Happy and an acceptable range of facial feature configurations.
Figure 3: Network responses to the Surprise $\rightarrow$ Happy transition for subject JJ. This figure suggests the existence of a complex transition space between categories, with the prediction of an intermediate perception of Fear as the face changes expression between Surprise and Happiness. Moreover, the middle region suggests a substantial break between the often correlated responses of Anger, Disgust and Sadness.
Subject “MF”: Morph from Happy to Sad

Figure 4: Network responses to the Happiness → Sadness transition for subject MF. This figure shows significant correlation between the Surprise and Fear networks, along with some lesser correlation between Disgust, Anger and Sadness. Indeed, Anger should here be taken as almost constant. Note the early onset of high activation for Sadness, due possibly to the similarity between Sadness and what Ekman and Friesen regarded as a Neutral category.
Figure 5: Network responses to the Happiness $\rightarrow$ Anger transition for subject JM. Here we observe behaviour qualitatively similar to the previous figure, with an intermediate Sadness peak and moderate correlation between Anger and Disgust. Note the negligible response of the Surprise network.

Figure 6: Network responses to the Anger $\rightarrow$ Surprise transition for subject JM. This set of plots shows again an intermediate peak for Sadness, but differs from the previous figure in preserving qualitative similarity between Sadness, Anger and Disgust.
Figure 7: Network responses to the Anger → Surprise transition for subject EM. This transition is significant in showing a breaking of the nexus between Fear and Surprise, with Fear almost constant and, for this subject at least, correlated closely with Happy. Note the presence of the usual correlation between Anger and Disgust, and the curious phenomenon that the Anger network shows some increasing response as the transition is completed into Surprise.

Figure 8: Network responses to the Happy → Disgust transition for subject GS. Note the bunching of responses for the Anger, Disgust and Sadness networks.
Table 1: Network responses for the unseen subject NR, using judge networks trained for 500 epochs. Network categories run horizontally (Happiness, Sadness, Fear, Anger, Surprise, Disgust) and images vertically, the numbers being those from the Ekman photographs. Each numeric column refers to a category network trained on all other images. Note the correct detection of each category and the strong responses for Disgust (as ‘perceived’ by the Anger network), and Fear (Surprise).

<table>
<thead>
<tr>
<th>NR</th>
<th>hp</th>
<th>sd</th>
<th>fr</th>
<th>ag</th>
<th>sr</th>
<th>ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>66: hp</td>
<td>0.60</td>
<td>-0.09</td>
<td>0.26</td>
<td>0.26</td>
<td>0.56</td>
<td>0.16</td>
</tr>
<tr>
<td>67: sd</td>
<td>0.30</td>
<td>0.72</td>
<td>0.36</td>
<td>0.54</td>
<td>0.14</td>
<td>0.34</td>
</tr>
<tr>
<td>68: fr</td>
<td>0.32</td>
<td>0.17</td>
<td>0.67</td>
<td>0.25</td>
<td>0.50</td>
<td>0.15</td>
</tr>
<tr>
<td>69: ag</td>
<td>0.50</td>
<td>0.57</td>
<td>0.13</td>
<td>0.83</td>
<td>-0.15</td>
<td>0.55</td>
</tr>
<tr>
<td>70: sr</td>
<td>0.27</td>
<td>0.05</td>
<td>0.46</td>
<td>0.38</td>
<td>0.62</td>
<td>0.45</td>
</tr>
<tr>
<td>71: ds</td>
<td>0.49</td>
<td>0.44</td>
<td>0.14</td>
<td>0.77</td>
<td>-0.30</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 2: Network responses for the unseen subject JJ, after training for 500 epochs. As before, each numeric column refers to a category network trained on all other images. Note the strong generalisation (see the elements of the main diagonal) and the WTA judgment that the ‘angry’ image should be perceived as ‘disgusted’.

<table>
<thead>
<tr>
<th>JJ</th>
<th>hp</th>
<th>sd</th>
<th>fr</th>
<th>hp</th>
<th>sd</th>
<th>sr</th>
</tr>
</thead>
<tbody>
<tr>
<td>34: hp</td>
<td>0.62</td>
<td>0.16</td>
<td>0.50</td>
<td>0.38</td>
<td>0.29</td>
<td>0.42</td>
</tr>
<tr>
<td>36: sd</td>
<td>0.25</td>
<td>0.47</td>
<td>0.53</td>
<td>0.39</td>
<td>0.28</td>
<td>0.50</td>
</tr>
<tr>
<td>37: fr</td>
<td>0.25</td>
<td>0.35</td>
<td>0.73</td>
<td>0.30</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td>38: ag</td>
<td>0.13</td>
<td>0.40</td>
<td>0.31</td>
<td>0.50</td>
<td>0.39</td>
<td>0.59</td>
</tr>
<tr>
<td>39: sr</td>
<td>0.18</td>
<td>0.31</td>
<td>0.57</td>
<td>0.24</td>
<td>0.73</td>
<td>0.39</td>
</tr>
<tr>
<td>40: ds</td>
<td>0.52</td>
<td>0.33</td>
<td>0.32</td>
<td>0.44</td>
<td>0.17</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Table 3: Network responses for the unseen subjects EM and GS, after network training for 300 epochs. While strong generalisation performance is indicated for most categories, these judge networks also show the common difficulty in ‘perceiving’ anger, disgust and sadness, with comparable activations being observed across the three categories for appropriate stimuli. The confusion between anger and disgust is especially pronounced, and is consistent with the human subject literature. Unusually, these examples also exhibit an high degree of similarity between the three stated categories and fear, although this may disappear with further training. The activation of the Happy network for image 14 provides further evidence that this category is more cleanly defined than some of the others.
<table>
<thead>
<tr>
<th>MF</th>
<th>Hp</th>
<th>Sd</th>
<th>Fr</th>
<th>Ag</th>
<th>Sr</th>
<th>Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 : Hp</td>
<td>0.75</td>
<td>0.15</td>
<td>0.13</td>
<td>−0.03</td>
<td>0.20</td>
<td>0.23</td>
</tr>
<tr>
<td>49 : Sd</td>
<td>0.43</td>
<td>0.64</td>
<td>0.43</td>
<td>0.63</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>50 : Fr</td>
<td>0.33</td>
<td>0.32</td>
<td>0.67</td>
<td>0.38</td>
<td>0.37</td>
<td>−0.17</td>
</tr>
<tr>
<td>53 : Ag</td>
<td>0.42</td>
<td>0.40</td>
<td>0.31</td>
<td>0.47</td>
<td>0.43</td>
<td>0.55</td>
</tr>
<tr>
<td>54 : Sr</td>
<td>0.26</td>
<td>0.50</td>
<td>0.84</td>
<td>0.51</td>
<td>0.64</td>
<td>0.15</td>
</tr>
<tr>
<td>55 : Ds</td>
<td>0.10</td>
<td>0.09</td>
<td>0.01</td>
<td>0.28</td>
<td>0.03</td>
<td>0.64</td>
</tr>
<tr>
<td>MO</td>
<td>Hp</td>
<td>Sd</td>
<td>Fr</td>
<td>Ag</td>
<td>Sr</td>
<td>Ds</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>57 : Hp</td>
<td>0.91</td>
<td>0.23</td>
<td>0.19</td>
<td>0.18</td>
<td>−0.02</td>
<td>0.39</td>
</tr>
<tr>
<td>58 : Sd</td>
<td>0.28</td>
<td>0.56</td>
<td>0.30</td>
<td>0.40</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>60 : Fr</td>
<td>0.29</td>
<td>0.14</td>
<td>0.96</td>
<td>0.22</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>61 : Ag</td>
<td>0.25</td>
<td>0.70</td>
<td>0.04</td>
<td>0.67</td>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>63 : Sr</td>
<td>0.06</td>
<td>0.20</td>
<td>0.46</td>
<td>0.11</td>
<td>0.63</td>
<td>0.22</td>
</tr>
<tr>
<td>64 : Ds</td>
<td>0.39</td>
<td>0.48</td>
<td>0.29</td>
<td>0.60</td>
<td>0.28</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 4: Network responses for the unseen subjects MF and MO after training for 300 epochs. These results provide very strong indications of the classical confusions between fear and surprise (note in particular the Fear response to image 54, and the Anger response to image 64). The results are also noteworthy for the extraordinary prototypicality of the MO Happy and Fear images (respectively image number 57 with activation 0.91, and image number 60 with activation 0.96).

<table>
<thead>
<tr>
<th>JM</th>
<th>Hp</th>
<th>Sd</th>
<th>Ag</th>
<th>Sr</th>
<th>Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>42 : Hp</td>
<td>1.14</td>
<td>0.14</td>
<td>0.13</td>
<td>0.05</td>
<td>0.27</td>
</tr>
<tr>
<td>43 : Sd</td>
<td>0.53</td>
<td>0.72</td>
<td>0.62</td>
<td>0.27</td>
<td>0.68</td>
</tr>
<tr>
<td>44 : Ag</td>
<td>0.29</td>
<td>0.57</td>
<td>0.76</td>
<td>0.28</td>
<td>0.70</td>
</tr>
<tr>
<td>45 : Sr</td>
<td>−0.11</td>
<td>0.10</td>
<td>0.06</td>
<td>0.84</td>
<td>−0.05</td>
</tr>
<tr>
<td>46 : Ds</td>
<td>0.24</td>
<td>0.42</td>
<td>0.45</td>
<td>0.03</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 5: Network responses for the unseen subject JM, after training for 300 epochs. Results for subject JM provide strong generalisation performance along with the usual similarity between Anger, Disgust and Sadness.
Acknowledgements

The authors would like to acknowledge helpful discussions with Claudia Brugman, and assistance with image processing from Sylvia Wyllie.
References


