A Receptive Field Neural Network which Learns to Describe Facial Expressions

James M. Hogan and **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 for denotata (here a range of facial images known to be associated with terms for human emotion) through the use of a receptive field based artificial neural network model. The network is trained upon images derived from the Ekman and Friesen "Pictures of Facial Affect" database, and is subsequently able to successfully generalise to images of unseen subjects. By using digital morphing techniques to produce intermediate frames between the existing stills, we predict that the space of transitions between denotata is potentially complex, and that such denotata may have only a limited role in the acquisition of more complex emotional terms.

1 Introduction

Labov's seminal work of 1973 on the fuzzy boundaries of denotational categories showed that denotation (defined here as the association of a linguistic label with a set of possible referents) is subject to the indeterminacy found in other areas of human categorisation (cf. (Mervis and Rosch, 1981)). His well- known CUP research shows that small variations in the features underlying denotata of lexical categories will contribute to a gestalt judgement of greater or lesser – equivalently, more or less prototypical – membership in a category that can be named by a lexical item. Such denotata inhabit a space akin to a spectrum, in which gradual differences give way at some point to categorical distinctions. It has been the work of linguistic semantics since then to lay out the complex ways in which human categorisation is associated with the application of linguistic labels (see also (Lakoff, 1987) (Allan, 1986) and (Taylor, 1996)).

This work uses an artificial neural network model to explore this phenomenon in the judgment of facial expressions as denotata of terms such as "Happy", "Sad", "Angry", "Fearful", "Surprised" and "Disgusted", the categories established by (Ekman and Friesen, 1975) in their famous "Pictures of Facial Affect" study. Though we do not claim that we are getting at the sense of each term (these being defined in relation to the underlying emotional state), we show that important aspects of characteristic facial expressions (which we may here consider the denotational range) can be abstracted into schematic images, which serve as scene-like inputs to be associated with particular descriptions. The present project is closely allied with the Berkeley Lzero project (Feldman et. al., 1990) in addressing the integration of visual and linguistic information through a learning task in which a description is associated with a scene. Here, as in Lzero, our system must determine whether a particular description is true of the accompanying image.

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. This phenomenon is further investigated through the use of digital morphing to create intermediate frames between the existing images.

This paper is organised as follows. Section 1.1 provides a summary of the relevant developmental literature, while section 2 discusses the network model and the receptive field units employed for each network. Details of the image preparation and learning task are given in sec-

* This work was funded by a grant from the Australian Research Council.

tion 3, and we conclude with discussion of the results and conclusions in sections 4 and 5.

1.1 Literature

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 recognition of a familiar subject. 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 (Johnson and Morton, 1991), (Hainline, 1978), (Tanaka and Farah, 1993), showing a marked preference for tracking facial stimuli as early as the first hour of life, suggesting some innate preference for conspecifics, later refined to the infant's caregiver once visual acuity has developed sufficiently. This preference continues through the first month of life, subsequently declining, only to re-emerge around a month later as a result of the maturation of primary visual cortex (Johnson and Morton, 1991). 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 (Johnson and Morton, 1991) – through, say, a minor adjustment to a line-drawn mouth. Hainline (1978), reports further development of scanning from 7 weeks through to the establishment of an integrated percept by the end of the first year of life.

Studies of changes in performance on facial expression tasks with age are limited, and variations in experimental approach make it difficult to trace development of face recognition through to adulthood, although in the latter case there is evidence of a near-monotonic improvement in performance from around five years to maturity (Chung and Thomson, 1995). It is clear however (Murray and Stanley, 1980), that by age 4-6, children may be trained to associate verbal descriptions (such as "happy") with line drawings of appropriate facial expressions, and that this capacity develops over time, being established well before adolescence (Dimitrovsky et. al. 1998). 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 (1995) postulate that improved recognition performance may be the result of memorisation of a greater number of facial features and feature relations.

2 The Network Model

The networks used in this study employ the traditional architecture of Radial Basis Function (RBF) networks, with a substantial hidden layer fully connected to a single output 'concept' unit. Significantly for the present task, RBF units are constrained to respond according to the (Euclidean) separation between the input vector and the unit centre, weighted according to the spatial Gaussian of equation 1. Influence thus decays strongly as separation increases, but remains strong within some neighbourhood of the centre.

More formally, given an input vector $x \in R^n$, a Gaussian RBF unit centred upon $c \in R^n$ will have activation:

$$G(x; T, c) = \exp(-||T(x - c)||_2^2), \qquad (1)$$

where $T \in R^{n \times n}$ is the *transformation matrix*, which determines the shape of the unit's receptive field. Usually T is chosen to be diagonal:

$$T = diag(1/r_1, \ldots, 1/r_n), \qquad (2)$$

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.

2.1 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, we have employed trainable Gaussian *domain response* units (Hogan, in preparation) as a mechanism for incorporating the influence of an entire region.

Consider an image over a bounded domain $D \subset R^2$ with normalised intensity level:[1]

$$I(x) : D \to [0, 1], \qquad (3)$$

[1]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 formally defined by:

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

where

$$< G >= \int_D G(x;T,c) dx, \quad (5)$$

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 rewritten:

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

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 of (Geva, 1992), 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 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. These pixel maps were 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 are still 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.

Due to the paucity of data, the nets are trained on a leave-one-subject out approach, in which *all* of the images of a single Ekman subject are reserved for testing. For each expression category, the 'judge' network is chosen as the best of eight training runs. An overall category judgment may be made through a winner-take-all across the networks.

4 Results and Discussion

Raw generalisation results are shown for unseen subjects NR and JJ in tables 1 and 2 below. Each judge network produces its maximal response for the set of images of a *single* subject on the image corresponding to its nominated category. However, this need not be sufficient to win the WTA competition, and significant responses may be observed from competing networks. Moreover, some of these confusions may occur more frequently than others, in particular those between "Surprised" and "Fearful" and "Angry" and "Disgusted", in accord with Ekman's human subjects. In addition, we observe consistent difficulty with images of the "Sad" category, there being significant responses from the "Angry", "Disgust" and "Fear" networks. This is reminiscent of the confusion Labov reports among subjects in attaching linguistic labels to the pictures he constructed which non-deterministically combined features of cups, mugs and bowls.

For each of the test images, digital morphing software was used to produce a series of 16 intermediate frames between the well-classified Ekman stills, with 48 morphing control points chosen manually. The linear interpolation of distorting control points and linear pixel value crossfade results in a natural and gradual transition between expressions. 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 monotonically decreasing and increasing activation levels. However the responses of the remaining categories suggest a perceptually confusing trajectory may be followed as the transition takes place. In the results for unseen subject GS shown in table 3, for example, transitions involving "Surprised" may frequently produce intermediate frames strongly characterised as "Fearful", and similar confusions may appear for "Angry" and "Sad". Significantly, the set of transitions suggests that our networks encode the relevant features within as few as four dimensions.

5 Conclusions

Our results require severe limitations on any hypothesis that transitions in facial expression denotata may be directly identified with transitions in the associated emotional state. Similarly, we note that the success of our low complexity model, and the correlation between its predictions and the cited developmental literature, suggest that a low dimensional encoding may be employed by children in representing the less difficult Ekman categories, and that a more complex representation is necessary for judgment between "Disgusted", "Fearful", "Sad" and "Angry". The necessity of such a representation – perhaps involving a greater number of features as found by (Ekman and Friesen, 1975), or as postulated in the recognition literature (Chung and Thomson, 1995) – and its late appearance, strongly suggests that expression denotata may have only a limited role in the acquisition of the semantics of more complex emotional terms.

Acknowledgements: The authors would like to acknowledge helpful discussions with Claudia Brugman.

References

K. Allan. 1986. *Linguistic Meaning.* Cambridge University Press.

M.S Chung and D.M. Thomson. 1995. Development of face recognition. *Br-J-Psychol.*

P. Ekman and W. Friesen. 1975. *Unmasking the Face.* Prentice Hall.

S. Geva. 1992. *Exponential response artificial neurons for pattern classification and function approximation.* Ph.D. thesis, Queensland University of Technology.

L. Hainline. 1978. Developmental changes in visual scanning of face and nonface patterns by infants. *Journal-of-Experimental-Child-Psychology.*

J.M. Hogan. in preparation. *Recruitment Learning in Randomly Connected Neural Networks.* Ph.D. thesis, Queensland University of Technology.

M.H. Johnson and J. Morton. 1991. *Biology and cognitive development: the case of face recognition.* Basil Blackwell.

G. Lakoff. 1987. *Women, Fire and Dangerous Things: what categories reveal about the mind.* U. Chicago Press.

C. Mervis and E. Rosch. 1981. Categorization of natural objects. *Ann. Rev. Psych.*

F.S. Murray and R.L. Stanley. 1980. Perceptual learning of cartoon faces by young children. *Bulletin-of-the-Psychonomic-Society.*

J. Taylor. 1996. *Linguistic Categorisation.* Cambridge University Press.

Image	ag	ds	fr	hp	sd	sr
69: ag	0.83	0.55	0.13	0.50	0.57	−0.15
71: ds	0.77	0.86	0.14	0.49	0.44	−0.30
68: fr	0.25	0.15	0.67	0.32	0.17	0.50
66: hp	0.26	0.16	0.26	0.60	−0.09	0.56
67: sd	0.54	0.34	0.36	0.30	0.72	0.14
70: sr	0.38	0.45	0.46	0.27	0.05	0.62

Table 1: *Network responses for the unseen subject NR. Network categories run horizontally (angry, disgusted, fearful, happy, sad, surprised) 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 confusion responses for disgusted (as 'perceived' by the angry network), and fear (surprised).*

Image	ag	ds	fr	hp	sd	sr
38: ag	0.50	0.59	0.31	0.13	0.40	0.39
40: ds	0.44	0.69	0.32	0.52	0.33	0.17
37: fr	0.30	0.39	0.73	0.25	0.35	0.37
34: hp	0.38	0.42	0.50	0.62	0.16	0.29
36: sd	0.39	0.50	0.53	0.25	0.47	0.28
39: sr	0.24	0.39	0.57	0.18	0.31	0.73

Table 2: *Network responses for the unseen subject JJ. As before, each numeric column refers to a category network trained on all other images. Note the WTA judgment that the 'angry' image should be perceived as 'disgusted'.*

Morph	ag	ds	fr	hp	sd	sr
agsr	0.10	0.08	0.27	0.07	0.18	0.09
hpag	0.25	0.36	−0.13	0.06	0.30	0.03
hpds	0.21	0.36	−0.14	0.17	0.31	0.00
sdsr	0.05	0.03	0.25	0.04	0.22	0.14
srds	0.07	0.09	0.19	0.12	0.24	0.07

Table 3: *Selected network responses for the* middle *frame of the morph indicated in the left most column. The end stills are those for the unseen subject GS. Note that for these selected sequences, the middle frame is best associated with some third category, suggesting non-linear transition pathways in the space of denotata.*