PREDICTING PHYSICAL FEATURES OF FACES FROM VERBAL DESCRIPTIONS: A COMPARISON OF NEURAL NETWORK AND LINEAR REGRESSION MODELS

Jeroen G.W. Raaijmakers
University of Amsterdam

Linear regression models and feedforward neural network models based on back-propagation were compared as to their ability to capture the relation between verbal descriptions of faces and physical measures from the same faces. Feedforward network models were very successful in fitting the original training exemplars but broke down under degraded input conditions, i.e., when noise was added to the input or when part of the input data was missing. It was shown that this is mainly due to the excessively large number of parameters in such models. It was also shown that this breakdown could be avoided by adding a small amount of random noise to the input, thereby preventing the model to tune itself to fine details of the input data. However, in all analyses the neural network models were significantly less successful in their generalization to new exemplars, exemplars that had not been seen during training. Thus, generalization seems to be an inherent problem for such feedforward neural network models.

Eyewitnesses to a crime often give a verbal description of the face of the perpetrator. Such a description is usually incomplete and the details are often incorrect. Yet such descriptions are important for judicial purposes and may be used in two ways. First, such data may be used to determine whether the description matches one of more persons from the police archives. Second, the descriptions are sometimes used to construct a composite picture or sketch of the criminal. For both purposes it would be useful to have a system that would help to transform the verbal description into a set of physical measures of face characteristics, e.g. measures for the relative size of the nose, the forehead, etc. Such a set of physical measures could then be used to search a database of faces to find the best matches or to construct a pictorial representation (e.g., to determine the best-matching Photo-fit).

Making better use of the information available in verbal descriptions of faces may increase the effectiveness of such procedures. There is some indication in the literature that verbal descriptions of faces provide more information than Photo-fit reconstructions (Christie & Ellis, 1981). Current procedures for the construction of such pictures are notoriously unreliable and not very effective. Ellis, Davies and Sheperd (1978) showed that Photo-fit composites made from memory were correctly matched in only 1 out of 8 cases by an independent group of subjects (chance level was 1 in 36). Even more disturbing is the finding that the quality of the composite was unaffected by whether or not the target face was or was not visible during the reconstruction (Ellis, Davies & Shepherd, 1978, Exp 2). A conclusion that one might be tempted to draw from such findings is that subjects are simply unable to generate a good description of faces from memory. Such a conclusion, however, does not seem to be correct. Other research (Ellis, Davies & Shepherd, 1978; Ellis, Shepherd & Davies, 1975; Laughery & Fowler, 1980; Shepherd & Ellis, 1973) shows that performance is good when a recognition measure instead of a recall measure is used. Moreover, subjects are twice as good in matching a verbal description to the target face as a composite based on Photo-fit (Christie & Ellis, 1981). Thus, verbal descriptions contain more relevant information than is captured using current techniques for the making of face composites.

The latter finding was the starting point for the current research. The finding shows that subjects are better capable to form a mental image of a face from a verbal description provided by an eyewitness than on the basis of a

I would like to thank John Shepherd of the University of Aberdeen for making the data available on which the present analyses were based, and C. Vijlbrief of the TNO Human Factors Research Institute for programming the neural network analyses. The research reported in this paper was supported by a grant from the Netherlands Ministry of Economic Affairs under the SPIN programme.

Requests for reprints should be sent to J.G.W. Raaijmakers, Department of psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: raaijmakers@psy.uva.nl.

Photo-fit composite that is constructed using such a description. Apparently there is more information in the descriptions than is currently used. Somehow subjects are able to extract this information from verbal descriptions in a way that is not captured by current techniques for constructing face composites. One important problem, however, is that verbal descriptions given by witnesses are often inaccurate and incomplete. A system that uses such descriptions should be quite robust to such factors. It was assumed that a prediction model based on neural networks might do better than more conventional prediction -models. Neural networks are generally thought to perform well under noisy conditions and/or missing data ('graceful degradation'). In addition, neural network or connectionist models have often been assumed to be especially useful in providing working solutions to difficult and poorly specified problems (e.g., Humphreys, 1993). As a result, connectionist models have been applied to many real-world problems, including face recognition (Aleksander, 1983; Stonham, 1986). Hence, in this analysis a set of descriptive measures of faces will be used to predict a set of physical measures for the same faces using a standard neural network approach. Since there is some discussion about whether such network models are really superior to more traditional methods, the network model will be compared to a linear regression model.

METHODS

Data were kindly made available by Dr. John Shepherd of the University of Aberdeen for a total of 350 faces. The data were a sample from a larger database (Shepherd, 1986) and consisted of descriptions and physical measures from the faces of male persons. The database from which these data were obtained was constructed in such a way that the faces were representative in terms of age distribution and the presence of features such as beards, mustaches and glasses for the population of men that get into contact with the police (for more details, see Shepherd, 1986).

For each face, data were available for a number of physical measures as well as subjective descriptions. Appendix A lists the physical measures that were used. These 30 measures were based on previous research by Jones, Hirschberg, Rothman and Malpass (1976) and Shepherd (1986). The technique that was used, started out from the coordinates of 37 points which were subsequently transformed to the length and area measurements described in Appendix A.

Appendix B lists the subjective rating scales that were used. Of these 50 measures, 4 were not used (viz., P43, P44, P45, en P47) because these variables had little or no variance in the set of 350 faces. In addition, variable P39 was omitted since it was completely dependent on three other variables: P40, P41, and P42 (P39=0 if all of these are 0 and P39=1 in all other cases). This leaves 45 descriptive measures that were used as input variables in the analyses. Each of these descriptive measures (with the exception of the data regarding height, weight and age) were obtained by averaging the data of 10 independent raters. All of the 30 available physical measures were used as output variables, i.e. the variables that the models should predict. The input data were rescaled to values between 0 and 1 (this is for convenience only and has no further consequences).

Of the 350 faces, 50 were reserved for a test of generalization performance. Of the remaining 300, one case had missing data and was omitted. Hence, a set of 299 faces was used to train the network (to estimate the optimal weights of the connections) and in the regression analyses.

Two models were used: a regular linear multiple regression model and a multiple-layer feedforward network. The network model was quite conventional. In addition to the 45 input units and the 30 output units (the predicted physical measures), there was a middle layer of so-called hidden units. All of the input units were connected to all of the hidden units and these were fully connected to all of the output units. There were two variants of the network: one model (NN-15) had 15 hidden units and the other model (NN-30) had 30 hidden units. In addition, the output variables were rescaled[1]. This is not trivial since the predicted values are equal to the output of a sigmoid transfer function. This function is approximately linear around 0.5. These analyses were performed on a UNIX workstation using the Rochester Connectionist Simulator (RCS version 4.1) developed by Goddard, Lynne and Mintz (1988).

We tried two versions: in one case the observed physical measures were rescaled

between 0 and 1, in the second case they were rescaled between 0.4 and 0.6. The latter variant implies that the transfer function for the output units is approximately linear (note that the transfer function for the hidden units is still nonlinear). Hence, 4 variants of a feedforward network were tried:

- NN1-15: 15 hidden units, scaling between 0 and 1,
- NN1-30: 30 hidden units, scaling between 0 and 1,
- NN2-15: 15 hidden units, scaling between .4 and .6,
- NN2-30: 30 hidden units, scaling between .4 and .6.

RESULTS

A comparison was made between two models for predicting the physical measures from the subjective ratings. The first model was a standard linear regression model. That is, each of the 30 physical measures (Y_k) was described as a linear function of the 45 ratings (X_j):

$$Y_k = b_{0k} + \sum b_{jk} X_j + \varepsilon_k \qquad (1)$$

The term ε_k gives the deviation of the predicted from the observed value of Y_k. The parameters b_{jk} (j=0,…, 45) were estimated in the conventional way using the method of least squares.

This model was compared to a multiple-layer feed-forward network model (see Rumelhart, Hinton & Williams, 1986). In such a model, the activation (predicted value) of an output unit (variable) is a nonlinear function of the activation of a number of hidden units whose activation is a nonlinear function of the input units:

$$H_j = F(\sum w_{ji} X_i + \alpha_j) \qquad (2)$$

$$T_k = F(\sum w_{kj} H_j + \alpha_k) \qquad (3)$$

where T_k is the predicted output, the w_{ji} are weight parameters, α_j and α_k are additive constants and F is the usual sigmoid transfer function:

$$F(x) = \frac{1}{1+\exp(-x)} \qquad (4)$$

The parameters of the network model (the weights) were estimated using the back-propagation or generalized delta rule proposed by Rumelhart, Hinton and Williams (1986). In this procedure a method similar to the steepest descent method is used to minimize the deviation between the observed (Y_k) and predicted (T_k) output. During the training phase, the set of 299 faces was repeatedly presented to the network. Once the network had reached an equilibrium point (had more or less converged), a new cycle was started using a smaller learning parameter (delta). This continued until no further improvement was possible.

In the analysis of the results four aspects were considered. First, it was investigated how well the various models performed for the data that were used in the training phase ("familiar faces", the data that were used to estimate the parameters of the models). Second, an analysis was made of the performance when random error is added to the input data (simulating unreliability). Third, the performance under conditions of missing data was determined. Finally, for each of the models it was investigated how well they performed when presented with new faces.

PERFORMANCE FOR KNOWN FACES

The measure that was used for the evaluation of the goodness-of-fit was based on the squared differences between observed and predicted values in comparison to the observed variance. The measure, R^2, is equal to the mean percentage of the variance that can be explained by the model and is comparable to a squared product-moment correlation coefficient:

$$R^2 = 100 \times \left[1 - \sum\sum (Y_{jk} - T_{jk})^2 / \sum\sum (Y_{jk} - \bar{Y}_k)^2\right] \qquad (5)$$

where Y_{jk} is the observed value on variable Y_k for face j, T_{jk} the corresponding predicted value and $\bar{Y}_k$ the average observed value. In addition the results were evaluated in terms of the average Euclidean distance, D, between the vector of the observed and predicted output variables:

$$D = \sum \left[\sqrt{\sum (Y_{jk} - T_{jk})^2}\right] / N \qquad (6)$$

where N is the number of faces in the data set. D is equivalent to the measure that is minimized to obtain the parameter estimates for both the regression model and the back-propagation model. It should be noted that this measure is dependent on the scaling of the output variables and hence not comparable between the two scaling methods.

The linear regression model gave a reasonably good prediction R^2=63.5 (the same for both scaling variants). For the distance measure *D* we found: *D*=553 for scaling between 0 and 1 and *D*=110.7 for scaling between 0.4 and 0.6. The observed solutions for the network models however gave a better fit. For NN1-15 we found: R^2=70.0 (*D*=504) and for NN1-30: R^2=74.1 (*D*=470). The fact that R^2 and *D* for NN1-30 are superior to those of NN1-15 is understandable because the NN1-30 model has more free parameters. A similar difference between the models with 30 and 15 hidden units was also observed for the output scaling between 0.4 and 0.6. The NN2 models gave a lesser fit, however, and were only marginally better than the corresponding linear model. The NN2-15 model gave R^2=63.9 (*D*=110.6) and the NN2-30 gave R^2=69.9 (*D*=101.1).

The most important conclusion is that the network model (especially the NN1-30 variant) does much better than the regression model. This is according to our expectation since the network model is a more general model (it handles both nonlinear as well as linear relations).

Insert Figures 1 and 1 about here

We also looked at how well such a system could be used to retrieve matching faces from a database. This is of course a problem that is very relevant for practical applications (e.g. automated database systems for crime investigations). For each face, the set of predicted physical measures was compared to each of the 299 faces. As a measure of similarity (or rather dissimilarity) we used the Euclidean distance between the vector of observed values and the vector of predicted values. We then computed the rank *n* for the target face. Figures 1 and 2 give the probability that the target face is among the *n* best-matching faces. Once again, the network model and in particular the NN1-30 model does better than the linear model.

PERFORMANCE UNDER "NOISE"

Traditionally, one of the reasons for advocating the use of neural networks has been their tolerance to moderate amounts of error in the input data (this property is often referred to as 'graceful degradation'). The claimed superiority of network models in the handling of error was also the primary reason for conducting the present analyses (verbal descriptions of faces may be assumed to be especially error-prone). In order to test this aspect, we added a varying amount of uniformly distributed random error to the input data. The amount was varied over a large range, viz. between .05 and .65 (remember that the input data were rescaled to lie between 0 and 1). More precisely, for a range of, say, .25 a random number was chosen between -.125 and +.125 and this was added to the original value. The new input values were confined to the original range, i.e. when adding noise would lead to a value of -.2, this value was set to 0. These new values were then used to compute predicted output values. Figures 3 and 4 give the results.

Insert Figures 3 and 4 about here

Contrary to our expectations, the network models do not perform very well under these conditions of noisy input. The traditional linear model appears to be much less sensitive to such disturbances than the network model. Moreover, there seems to be a negative relation between performance for the original data and the performance under noise: The NN1-30 model did best for the original data but shows the greatest decline when random error is added to the input. One reason for this lack of robustness might be that the superior performance of the network model for the original data is largely due to the extremely large number of parameters (weights) in this model: the network model with 30 hidden units has 2310 parameters, whereas the linear model has 'only' 1380 free parameters. The difference between the NN1 and the NN2 models suggests that the greater flexibility of the NN1 models resulting from the nonlinearity for the output units leads to a better fit for the original data but leads to worse performance under noise.

PERFORMANCE WITH MISSING DATA

In this analysis we investigated the effects of missing data. This is another aspect for which network models are usually thought to be superior to more traditional models. In the present analyses, missing values were replaced by the mean value for that parameter (this is a

rather crude and not very intelligent way of handling missing data; however, it suffices for this exploratory study). In this manner 20% of the input data were replaced.

Somewhat surprisingly, in this analysis the network model again did not perform as well as the linear model. For scaling between 0 and 1, the results (in terms of the percentage of variance explained) were: linear: R^2=50.5; NN1-15: R^2=37.9; NN1-30: R^2=36.5. For scaling between .4 and .6 the results were: linear: R^2=50.5; NN2-15: R^2=51.8; NN2-30: R^2=50.8. Once again, it is the NN1-30 model that performs worst. And once again, there seems to be a negative relation between performance for the original data and the performance with missing values.

GENERALIZATION TO NEW FACES

All of the models were also tested on a new set of 50 faces that had not been used previously, i.e. these data were not used during the training phase. Once again, the performance of the linear model was better than that of the network model (especially NN1-30). For scaling between 0 and 1 the results were: linear: R^2=42.6; NN1-15: R^2=16.7; NN1-30: R^2=4.0. For scaling between .4 and .6 the results were: linear: R^2=42.9; NN2-15: R^2=38.1; NN2-30: R^2=23.5. Thus, the network model does not generalize nearly as well as the linear model. Note also that the performance of the NN1-30 model on this test of generalization is not just inferior to the linear model but also quite poor in absolute terms.

DISCUSSION

There are two main conclusions. First, the use of a neural network model based on back-propagation leads to better predictions for the original data compared to a conventional linear regression model. However, contrary to expectations, such a model did not lead to better performance under more critical conditions, i.e. when noise was added to the input or when tested with missing data. In addition, the back-propagation model generalizes less well than the linear model to a new set of input data. Although it has sometimes been claimed that the back-propagation model has excellent generalization properties (e.g. NETtalk, see Sejnowski & Rosenberg, 1987), it has been previously demonstrated that the generalization capabilities of back-propagation can be considerably improved by imposing a hidden layer bottleneck, i.e. a layer with relatively fewer units than previous layers (see Kruschke, 1989a,b). This suggests that the generalization problems might be related to the versatility of the back-propagation model.

In order to better understand why the back-propagation model did not perform as well as the linear model except on the original data, it is helpful to consider the number of parameters that were estimated for each of these models. For the linear model 1380 free parameters (=30*46) were estimated from the data, for the NN15 models the number of estimated parameters was 1170 (=15*46 + 30*16), while the NN30 models had 2310 estimated parameters (=30*46 + 30*31). Such a large number of parameters relative to the number of data points may lead to inflated estimates of the extent to which the model accurately accounts for the data. As was recently shown by Myung (1998), there is a general inverse relation between model complexity and generalizability. A model with many free parameters will fit a given data set better but will generalize more poorly to new data sets.

Similar problems are of course well known in the standard regression analysis and have led to the use of adjusted R^2 coefficients. In the present analyses there is indeed a negative relationship between the goodness-of-fit to the original data and the extent to which the model generalizes adequately to more demanding situations. Apparently, the relatively good fit to the training data is obtained by adjustment of the model parameters to minor and case-specific details of the data. Hence, the more freedom a model has to adjust its predictions, the greater the danger of over-fitting the data.

This aspect may explain why the NN-30 models did more poorly on the generalization test than the linear model. The above argument may be extended to account for the performance of the NN1-15 model relative to the linear model. As shown above, there is not much difference in the number of parameters in these models. Why then does the NN1-15 model perform worse on the generalization test than the linear model? We believe that this is due to the fact that the NN1-15 model is a more

flexible model than the linear model. That is, there is no restriction in the NN1-15 model to a particular function relating input to output.

The same reasoning might also account for the superior performance of the NN1-30 model compared to the NN2-30 model: in the latter case the transfer function that maps the hidden units to the output units is more or less restricted to a linear function (since the output values are restricted to the range 0.4-0.6). As in the analysis of Kruschke (1989b), this restriction appears to be beneficial to the generalization capabilities of the back-propagation model.

The rather poor performance of the back-propagation model in generalization is not a novel finding. Although often claimed otherwise, even the widely publicized NETtalk model (Sejnowski & Rosenberg, 1986) shows the same phenomenon. NETtalk has 203 input units, 80 hidden units and 26 output units, hence a total of 18426 parameters. NETtalk was trained on a set of 1024 words where it attained a score of 95% percent correct (for individual phonemes). When tested on a novel set of words, performance dropped to 78% correct (in terms of correct performance for whole words the results are even clearer: a drop from about 75% correct to about 15%). Although I am not aware of any data that looked at the performance of NETtalk with missing or noisy input, these results suggest that it is likely that that model will also not perform well on these tasks.

Hence we may conclude that the back-propagation model suffers from 'overfitting' and that its apparent superiority to more traditional regression techniques may be partly due to the extremely large number of parameters in such neural network models.

PERFORMANCE AFTER TRAINING WITH NOISY DATA

In order to get a more realistic estimate of the adequacy of the back-propagation model, some method has to be used to eliminate the problem that the model is simply capturing part of the error variance. One possibility is to train the model not on the original data but on sets of data obtained by adding a small amount of error variance to the original data. Thus, in the next analyses we trained the model not on the original data but on sets of data obtained by adding of uniformly distributed random error to the original data. To be more precise, on each training trial and for each input variable, a random number was chosen from a uniform distribution in the range (-.075, +.075). This number was then added to the original input value, again using the restriction that the resulting values should remain within the range [0,1]. This was done independently for each case (each face). Since new random values were used on each run through the whole sequence of faces, this procedure makes it impossible for the network model to adjust itself to minor variations of the input data. The resulting parameter estimates should then be more robust to noise and hopefully generalize better to new, not-seen-before, data. Since the problem was most evident in the NN1-models, these new analyses were only performed in for the NN1-15 and NN1-30 models (scaling of output units between 0 and 1).

For the linear regression model a similar procedure was used. However, since it is not possible to analyze unlimited numbers of input data, in this case ten data sets were generated using the method described above. The parameter estimates were obtained by submitting the combined data to one overall regression analysis. Hence, the resulting data set consisted of a total of 2990 cases.

After the parameter estimates had been obtained, the resulting models were applied to the original (error free) data. In addition, all of the analyses discussed previously were repeated (i.e., performance under noise and with missing data, and the generalization test).

As one might have expected, the results for the linear regression model were comparable to those obtained previously. For the original errorless data, the percentage of variance explained was R^2=63.0 (D =557). Interestingly, the network models still gave a better fit. For NN1-15 we found: R^2=67.1 (D=528) and for NN1-30: R^2=73.0 (D=478).

Insert Figures 5 and 6 about here

Next, we looked at how well the models could be used to retrieve matching faces from a database. As before, we computed for each set of predicted physical measures the similarity to each of the 299 faces and computed the rank for the target face. The results are shown in Figure 5 where the probability is given that the target face is among the n best-matching faces. Once again, the network model and in particular the

NN1-30 model does better than the linear model.

These results show that the addition of a small amount of random noise to the input has not seriously affected the ability of the models to correctly predict the physical measures from the verbal descriptions. However, the most important issue is whether such a procedure helps to eliminate the rather poor performance of the network model when the input is degraded. As before, we computed for all three models the ability to predict the output when a varying amount of random noise is added to the input. Figure 6 shows the percentage of explained variance of the output measures as a function of the amount of noise added. Clearly, the ability of the network models to cope with such degraded input has been greatly improved as can be seen by comparing Figure 6 to Figure 3. Interestingly, the regression model also performs better, especially when the amount of noise is relatively large. However, these results also show that under such conditions all three models converge to the same performance. Hence, these results show that the procedure that was used to obtain more robust parameter estimates was indeed successful but that there does not seem to be anything special about the ability of the network models to handle degraded input.

A similar conclusion is obtained from the analysis in which 20% of the input data were replaced by their mean values (simulating the effect of missing data). The results (again in terms of the percentage of variance explained) were: linear: R^2=51.5; NN1-15: R^2=51.0; NN1-30: R^2=52.6. All models perform about equally well and better (especially the network models) than in the original analysis.

How about the capability of the models to generalize to new stimuli not seen before? We tested the performance of the three models on the same set of 50 faces that was used in the previous test of generalization. The results were: linear: R^2=43.1; NN1-15: R^2=30.6; NN1-30: R^2=27.1. Thus, although the generalization performance of the network models is greatly improved, it is still inferior to that of the regression model. It should be noted that this is the only performance test that is not in some sense based on the original training data. Hence, it might be that the back-propagation model still has some generalization problems due to the fact that it is inherently more flexible. Whether or not this result is specific to the present example (faces may be a class of stimuli that are indeed best described by linear functions, see Abdi et al., 1995), is something that has to be clarified in future research.

GENERAL DISCUSSION AND CONCLUSIONS

The results reported in this paper clearly show that back-propagation models do not have a special ability to handle noisy data. Although these models were able to learn complex mappings between input and output data, this ability does not imply that such models will also perform well under more stringent conditions. Thus, in the present analyses, the network model broke down when noise was added to the input. This was especially true for the model with a relatively large number of hidden units. Although this model gave a superior fit when applied to the data on which it was originally trained, it performed quite poorly when random error was added to the input. It also performed worse than the regression model when part of the input data was replaced by missing data. Hence, contrary to what is frequently claimed, the back-propagation model does not seem to have a special ability for 'graceful degradation'.

We believe that the rather poor performance of the back-propagation model under degraded conditions is due to the flexibility and especially the extremely large number of parameters in these models. This enables such models to tune themselves to specific characteristics of individual training stimuli, characteristics that are not predictive for other exemplars within the same set. This property leads to very good predictions for the set of stimuli on which the model is trained (those stimuli that are used to estimate the parameters of the model) but also to the poor performance when such stimuli are slightly altered, for example by adding random noise.

In order to prevent this, either the size of the training set has to be increased or some other means must be used to force the model to focus on the most important relations between the input and output variables. One way to achieve this is to decrease the number of hidden units. Thus, we showed that reducing the number of hidden units from 30 to 15, decreases performance on the training exemplars but increases the performance when the input is degraded. An alternative method that seems to be more promising is to add a small amount of random noise to the input data.

As we showed in the final analyses, this does not affect the performance on the training set but greatly improves performance under degraded conditions. In the present case, the performance of all models (regression as well as back-propagation) converged to the same level as the amount of added noise was increased. Interestingly, all models benefited from this method, although the advantage was most significant for the back-propagation model with the larger number of hidden units.

The present finding that regression models perform at least as well as feed-forward neural network models, might be criticized as being specific for the current application. Thus, one might assume that the present application just happens to involve more or less linear relations between the input and output variables and hence it is not surprising that the linear regression model outperforms the neural network models. However, if this was indeed the case, then why should the neural network models perform better than the regression model on the original training set? It should also be noted that such back-propagation models have been claimed to be able to handle any type of relationship, hence they should not have any difficulties with a simple linear relationship.

A second interesting comparison between the regression model and the back-propagation model involves the extent to which the models capture the underlying relations so as to be able to generalize to new instances. It turns out that the regression model is definitely superior with respect to generalization. Interestingly, the rather poor generalization performance of the back-propagation model remains even when the model is trained on noisy input. Hence, whereas the poor performance on degraded input seems to be mainly due to the large number of parameters, the poor generalization seems to be due to other factors that cannot be so easily rectified.

All in all then, we arrive at the rather sobering conclusion that feedforward neural network models are not as successful as they seem to be when one only considers the extent to which the model is able to make correct predictions for the set of exemplars on which it was trained. Although such comparisons are not usually made, traditional regression models may in fact outperform the network models, at least when one takes into account such aspects as the ability to generalize to new data.

REFERENCES

Aleksander, I. (1983). Emergent intelligent properties of progressively structured pattern recognition nets. *Pattern Recognition Letters*, **1**, 375-384.

Christie, D.F.M. & Ellis, H.D. (1981). Photofit constructions versus verbal descriptions of faces. *Journal of Applied Psychology*, **66**, 358-363.

Ellis, H.D., Davies, G.M. & Shepherd, J.W. (1978). A critical examination of the Photofit system for recalling faces. *Ergonomics*, **21**, 297-307.

Ellis, H.D., Shepherd, J.W. & Davies, G.M. (1975). An investigation of the use of the Photofit technique for recalling faces. *British Journal of Psychology*, **66**, 29-37.

Goddard, N.G., Lynne, K.J. & Mintz, T. (1988) *Rochester Connectionist Simulator*. (Tech. Rep. 233). Computer Science Department, The University of Rochester, Rochester, New York.

Humphreys, G.W. (1993). Prospects for connectionism: Science and engineering. In A. Sloman et al. (Eds.), *Prospects for artificial inteligence*. IOS Press.

Jones, L.E., Hirschberg, N., Rothman, J. & Malpass, R.S. (1976). *The face atlas: Anthropometric, cosmetic and physiognomic measurements of 200 male faces*. Tech. Rep. 1, NSF GS-42801, Department of psychology, University of Illinois, Champaign, ILL.

Kruschke, J.K. (1989a). Creating local and distributed bottlenecks in hidden layers of back-propagation networks. In D. Touretzky et al. (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*. San Diego, CA: Morgan Kaufmann.

Kruschke, J.K. (1989b). Distributed bottlenecks for improved generalization in back-propagation networks. *Neural Networks*, **1**, 187-193.

Laughery, K.R. & Fowler, R.H. (1980). Sketch artist and Identikit procedures for recalling faces. *Journal of Applied Psychology*, **65**, 307-316.

Myung, I.J. (1998). Importance of complexity in model selection. *Journal of Mathematical Psychology*, in press.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Sejnowski, T.J. & Rosenberg, C.R. (1986). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145-68.

Shepherd, J.W. (1986). An interactive computer system for retrieving faces. In H.D. Ellis et al. (Eds.), *Aspects of face processing*. Dordrecht, The Netherlands: Martinus Nijhoff Publishers.

Shepherd, J.W. & Ellis, H.D. (1973). The effect of attractiveness on recognition memory for faces. *American Journal of Psychology*, **86**, 627-633.

Stonham, J. (1986). Practical face recognition and verification with WISARD. In H.D. Ellis et al. (Eds.), *Aspects of face processing*. The Hague: Martinus Nijhoff.

NOTES

[1] Rescaling of the output variables is necessary because the network model (i.e. the sigmoid transfer function) produces values between 0 and 1 at the output units.

APPENDIX A: PHYSICAL MEASURES

In the data collected by Shepherd (see Shepherd, 1986) photographs were used of a large number of faces (frontal view). For each face the exact (relative) location of 37 pre-defined points was measured (e.g., midpoint of hairline, midpoint of upper lip). On the basis of these points 30 derived measures were defined. These 30 measures were used in the present research as the to-be-predicted physical measures. They may be grouped in three categories:

DISTANCE MEASURES

M1	inter ocular distance
M2	forehead height
M3	nose length
M4	mouth width
M5	upper lip thickness
M6	lower lip thickness
M7	chin height
M8	face height
M9	face width at brow
M10	face width at cheek
M11	face width at mouth
M12	face width at chin
M13	eyebrow height
M14	eyebrow width
M15	eyebrow setting
M16	hair length
M17	nose width at bridge
M18	eye narrowness

AREA MEASURES

M19	face area
M20	hair area
M21	eye area
M22	chin area
M23	mouth area
M24	nose area

RATIO MEASURES

M25	eye:face ratio(M21/M19)
M26	mouth:face ratio(M23/M19)
M27	nose:face ratio(M24/M19)
M28	hair:face ratio(M20/M19)
M29	chin:face ratio(M22/M19)
M30	face height:width ratio (M8/M10)

APPENDIX B: RATING SCALES

The dataset that was available consisted of 50 ratings (means based of 10 raters). Most ratings (P1-P38) were obtained using 5-point rating scales. Parameters 39-47 were scored dichotomously. (0=no, 1=yes). The remaining 3 parameters (P48-P50) consisted of height, weight, and age. Of these 50 parameters 5 were not used in the analyses, viz. P39, P43, P44, P45, and P47.

SHAPE OF THE FACE
P1: short-long
P2: narrow-broad
P3: bony-fleshy

COMPLEXION
P4: fair-dark
P5: pale-florid
P6: unlined-lined
P7: clear-blemished

HAIR
P8: short-long
P9: tidy-untidy
P10: straight-curly
P11: bald-full head
P12: no grey-white
P13: black-brown-red-fair-blond

FOREHEAD
P14: low-high
P15: narrow-broad
P16: straight-sloping

EYEBROWS
P17. thin-thick
P18. straight-bent
P19. meet in the middle-set far apart
P20. low-high

EYES
P21. small-large
P22. narrowed-open
P23. close set-wide spaced
P24. deep set-protruding
P25. blue-grey-green-hazel-brown.

EARS
P26. small-large

NOSE
P27. small-large
P28. short-long
P29. narrow-broad
P30. concave-hooked
P31. small nostrils-large nostrils
P32. narrow tip-broad tip.

MOUTH
P33. small-large
P34. thin upper lip-thick upper lip
P35. thin lower lip-thick upper lip

CHIN
P36. small-large
P37. pointed-square
P38. receding-jutting

FACIAL HAIR
P39. none at all
P40. mustache
P41. sideburns
P42. beard

PHYSICAL PECULIARITIES.
P43. squint
P44. bags under eyes
P45. scars

ACCESSORIES
P46. glasses
P47. earring
OTHER
P48. height
P49. weight
P50. age

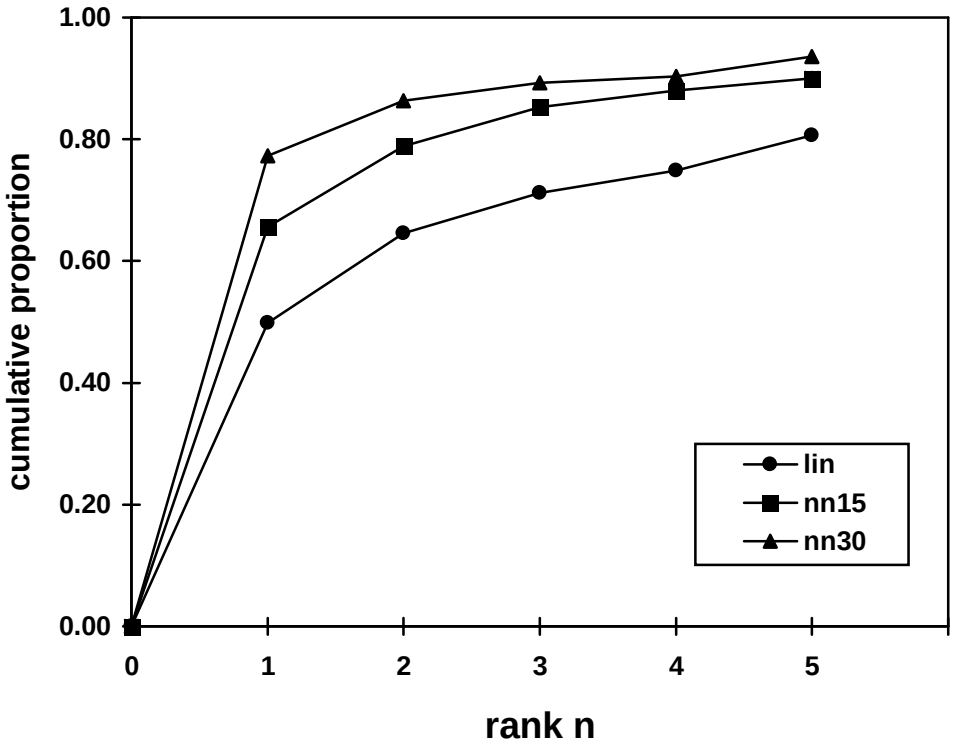

Figure 1. Probability that the target is among the *n* best-matching faces (scaling between 0 and 1).

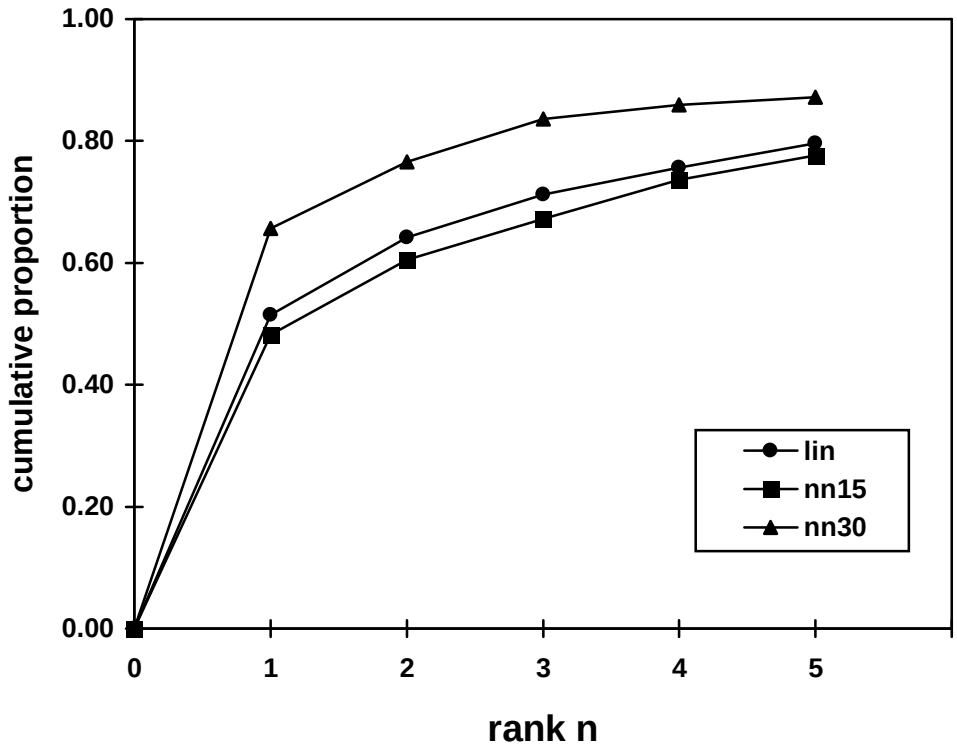

Figure 2. Probability that the target is among the n best-matching faces (scaling between 0.4 and 0.6).

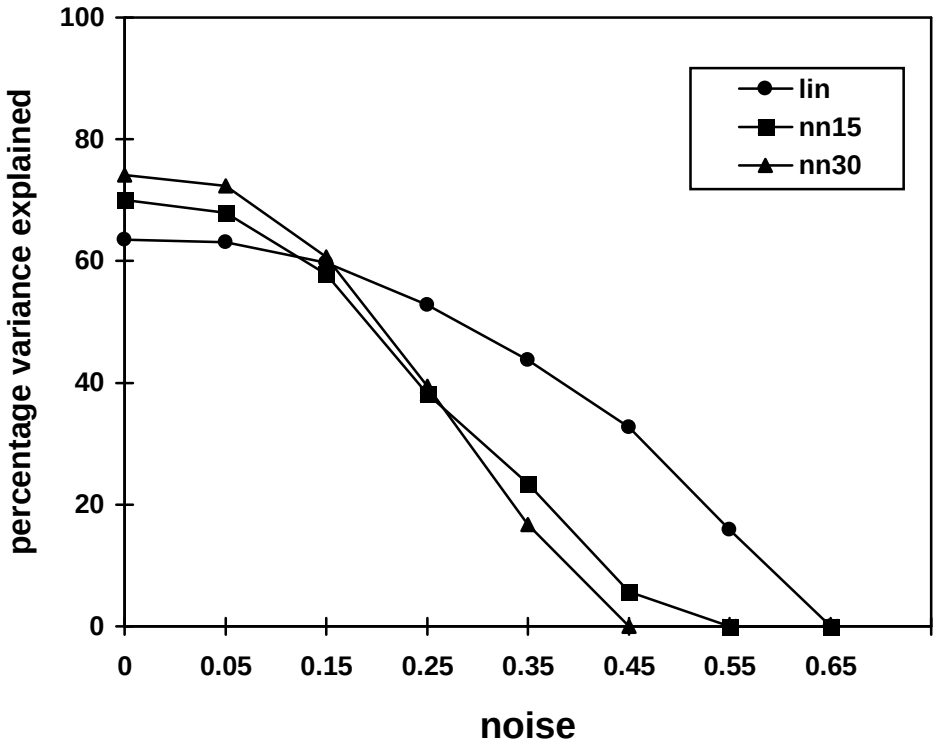

Figure 3. Percentage variance explained as a function of the amount of noise added to the input (scaling between 0 and 1).

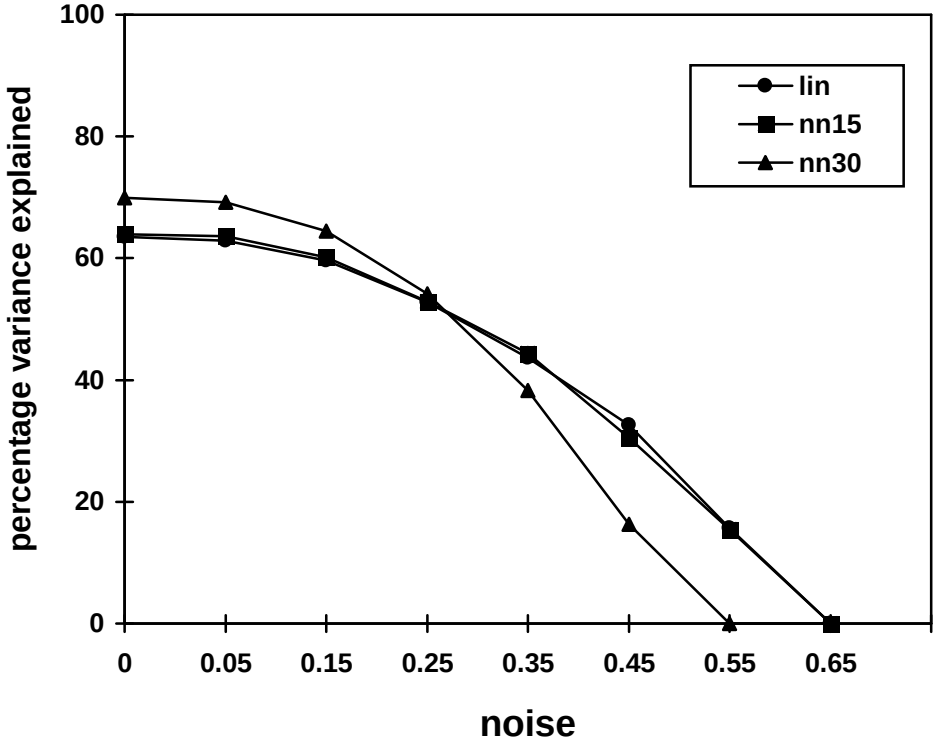

Figure 4 Percentage variance explained as a function of the amount of noise added to the input (scaling between 0.4 and 0.6).

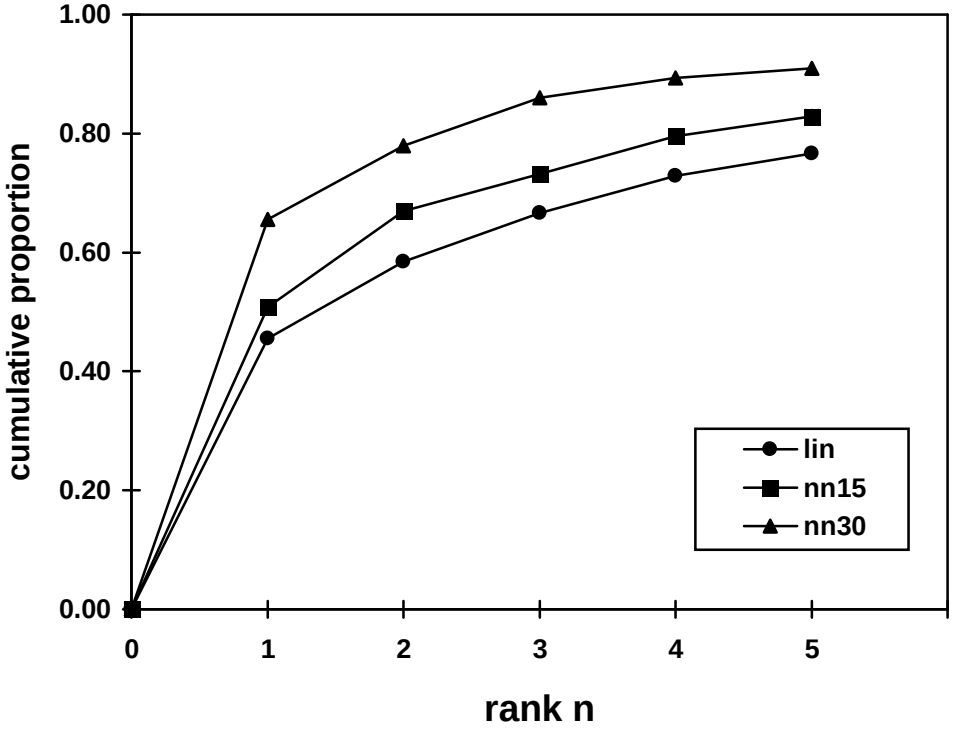

Figure 5 Probability that the target is among the *n* best-matching faces (training using noisy input).

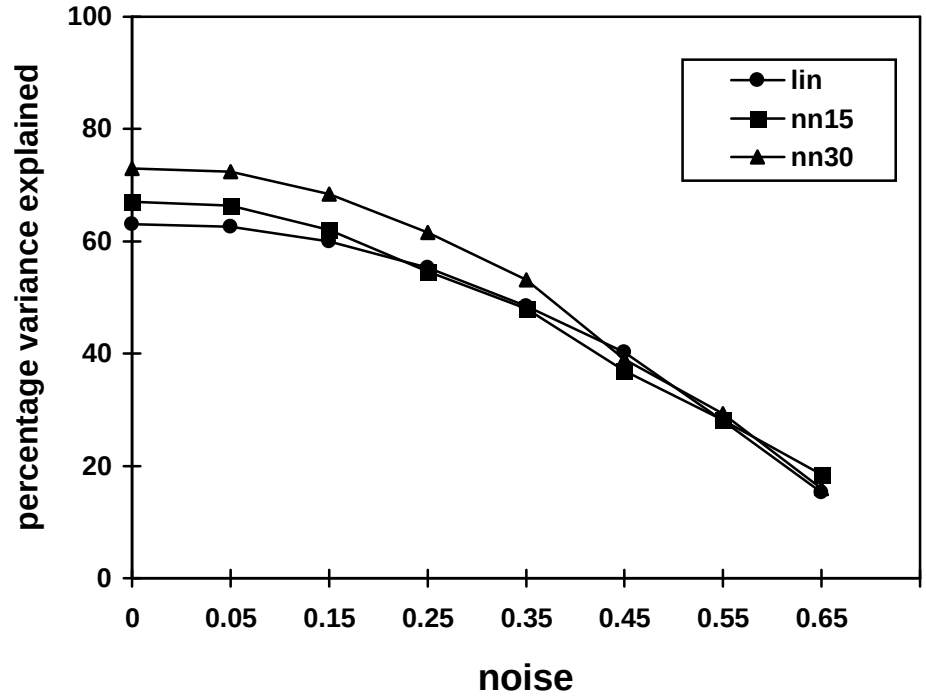

Figure 6 Percentage variance explained as a function of the amount of noise added to the input (training using noisy input).