# Artificial Neural Network Design and Parameter Optimization for Facial Expressions Recognition

Ammar A. Alzaydi

**Abstract**— This paper presents an Artificial Neural Network design and Neural Network parameter optimization for emotional recognition of classified facial expressions. The main goal in this paper is to teach computers to recognize three distinct human emotions from static images. Training and Testing dataset will be collected and a multilayer perceptron network will be built to implement an emotion classifier. Two excellent face databases are used to construct the training and testing datasets. Cross-validation techniques were used to compare the parameters of the Neural Network classifier and the types of activation functions. This paper shows that the performance of the designed Neural Network is very high at above 90% with around 60/40 ratio of training dataset size to test dataset size.

**Index Terms**— Artificial Neural Network, Design, Facial Expressions, Optimization, Parameter Optimization, Recognition

## 1 Introduction

As computers become ubiquitous in our every day lives, it is necessary to give them to ability to sense and appreciate human emotion. This paper seeks to tackle this Human Machine Interface problem using Artificial Neural Networks. The main goal in this paper is to teach computers to recognize three distinct human emotions from static images, which can be extracted from live video, such as a direct camera feed, to a robot/computer brain. Training and Testing dataset will be collected and a multilayer perceptron network will be built to implement an emotion classifier. The three emotions which are to be recognized are astonished, smiling, and calm. More emotion classes can be added in future work. Two excellent face databases, Yale and JAFFE, are used to construct the training and testing datasets [1], [2], [3]. Cross-validation techniques were used to compare the parameters of the Neural Network classifier such as the number of hidden nodes and the types of activation functions. From these tests, 25 hidden layer nodes were selected and used the hyperbolic tangent sigmoid activation functions for those nodes. A preprocessing step which consists of manual mouth alignment as well as synthetic illumination transformation during the training step is also devised to reduce network over-fitting and increase accuracy. This paper shows that the performance of the designed Neural Network is very high at above 90% with around 60/40 ratio of training dataset size to test dataset size.

Section 2 gives a quick review of relevant prior art regarding emotion or facial expression recognition using Neural Networks. Section 3 describes the approach from preprocessing to actual Neural Network classifier implementation. Section 4 gives empirical testing results from both the optimization step as well as the final evaluation step. Appendix A and B contains the image data used in the training and testing datasets.

• *Ammar A. Alzaydi: B.Sc., M.A.Sc., Ph.D. Student, Mechanical and Mechatronics Engineering. University of Waterloo, Waterloo, On., Canada. E-mail: aalzaydi@engmail.uwaterloo.ca*

## 2 Prior Art

One of the first studies into using a 3 layer Artificial Neural Network to determine emotion is described in [4]. In this paper, the author didn't perform any pre-processing such as alignment or illumination correction since their dataset didn't demand for it. Three regions of the face, forehead/eyebrows, eyes, and mouth/chin, were areas that fed directly into the input layer of the network. The results showed that the system generalized onto unseen data well only if the person was exposed to the network.

Another interesting method for expression recognition is proposed in [5]. In this paper, the authors propose to first perform a series of contour extraction followed by mathematical morphological operations to pre-process the face images. The result is then fed into a Neural Network classifier for emotion recognition. There result is very promising. However, by performing a serious of image processing steps, a lot of image assumption such as low-noise or sharp contrasted edges must hold in order for the algorithm to work.

Unsupervised methods have also been used to tackle emotion recognition [6]. In the method in [6], a self-organizing layer is inserted before the input layer. The self-organizing layer basically performs clustering and reduces the dimensionality of the input. The training is done using Hebbian learning and it is very similar to Principle Component Analysis. Specifically, it is equivalent to a localized version to Karhunen-Loève transform. While this method has been shown to have good results, it also needs a lot of sample images to compute the dimension reduction weights.

## 3 Implementation Details

The Neural Network classifier must transform its input in the form of grayscale images to the output in the form of a single number, with values matched to different emotions. This entire process is performed by preprocessing the raw image, propagate forward the activation of the input nodes all the way to the output, and finally assigning each image with an emotion based on the value of the output node.

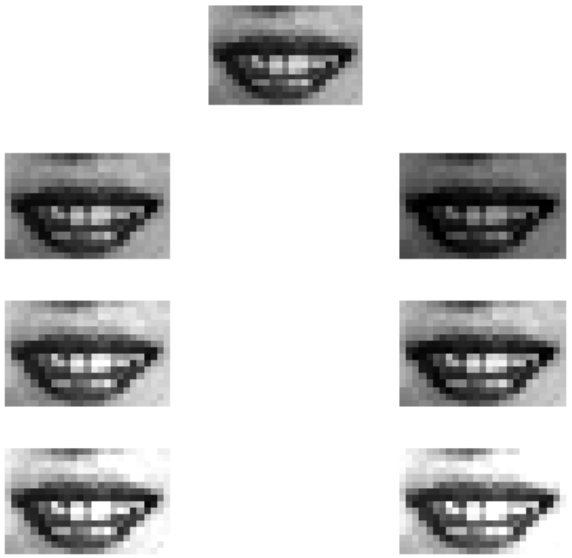


Fig. 1. Synthetic Lighting Transforms

### 3.1 Preprocessing

The problem of expression or identity recognition of the face is usually a combination of two sub-problems: alignment and classification [4], [5]. The alignment step seeks to align faces to a canonical grid such that the features of the face match exactly. For example, for faces of two different persons or different faces of the sample person at different times, it is necessary to match the exact location of key points of the two faces to facilitate matching and recognition. The solution to this problem in this paper's framework is to let the user manually click the left and right corners of each mouth. Using the location of each mouth, a 2D affine warp is performed to warp the two corners onto two set points inside a 16x25 pixels mouth image. The corresponding image pixel values in the original face image are sampled into the 16x25 pixels mouth image. Consequently, the normalized mouth image, which is 16x25 pixels in dimension, will contain 400 pixel values, each varying from 0 to 1.0. Each pixel is an input into the Neural Network classifier.

### 3.2 Illumination Invariance

Neural Networks trained with so little data and such large dimension weight space (400 input nodes) is doomed for over fitting when the hidden layer nodes increases. One simple way to alleviate this problem and also providing invariance to illumination is to artificially generate life-like mouths which are subjected to changing lighting conditions. Fig. 1, illustrates this technique. The top mouth image is the original image. The 6 images in Fig. 1 are images which underwent affine lighting transformation in the form of alpha*X + beta, where alpha and beta are uniform and Gaussian random variables, respectively.

The effect of the transformation is that it created realistic looking mouth under various lighting conditions. All images are entered into an enlarged training dataset for the Neural Network classifier. Therefore the classifier will basically see 6 times more points as it would have before. This addition reduces over-fitting and allows the classifier to perform well under varying lighting conditions. The recognition rate increased around 8.4% as a result of this step.

### 3.3 Neural Network Architecture

Artificial Neural Networks encompasses a rich array of architectural structures and application domains. One of the tasks that beset the designer is the choice of the initial Artificial Neural Network topology. The utility of the completed Artificial Neural Network design may be evaluated as optimized if the simplest implementation of an effective system is set forth.

Artificial Neural Network topologies may be either feedforward, recurrent or a hybrid of the two. Recurrent networks (e.g., Hopfield Artificial Neural Network) are best for, "any physical system whose dynamics in phase space is dominated by a substantial number of locally stable states to which it is attracted" [8]. Recurrent networks also tend to be more computationally expensive and storage intensive compared to feedforward systems. In contrast feedforward topologies have been successfully implemented for simplified classification problems. Given the tight constraints and sharp contrast in the classification categories of astonished, calm, and smiling for the mouth facial region, the feed-forward topology was chosen as the initial topology candidate for Artificial Neural Network construction.

The second step in establishing the Artificial Neural Network architecture is the choosing of the Artificial Neural Network type. Types range from Multi-Layered Perceptron, to Radial Basis Function Networks, to Kohonen Self-Organizing Networks. Kohonen Self-Organizing Networks types are not ideal for this application domain since a photographic training set with assessed emotions (i.e., targets) may be presented to the custom Artificial Neural Network *a priori*. Kohonen Self-Organizing Networks types are best suited for situations when targets are not known and unsupervised learning is employed to determine where the undiscovered classification boundaries lie. Next, Radial Basis Function Networks types may be suitable for this application but come with an additional level of undesired complexity. Radial Basis Function Networks systems are challenged with first performing an unsupervised training algorithm to determine the adaptation of centers and secondly performing a supervised training algorithm to tune the network connection weights. The balancing of these two steps of training introduces undesired complexity and potential for poor design. Multi-Layered Perceptron types have proven in past situations to provide adequate performance for simplified classification problems. As such, the authors choose Multi-Layered Perceptron as the initial type candidate for Artificial Neural Network construction.

### 3.4 Neural Network Parameters Optimization

The design of a feed-forward Multi-Layered Perceptron Artificial Neural Network is described by a number of internal parameters. The most basic specification of a Multi-Layered Perceptron Artificial Neural Network is the characterization of its input layer and output layer. This problem has been specified for design by the introduction of a 400 pixel dataset se-

lected (i.e., pre-processed) from the source image. The outputs of the classifier are the three emotional states of astonished, calm, and smiling. If a true versus false test is asked of the classifier for each of the emotions, then by extension three output nodes are required.

The number of hidden features of the Artificial Neural Network must also be determined. First, the number of layers interposing between the input and output layers must be specified. Once the number hidden layers are set, the number of hidden nodes within each layer must be set. Optimizing these parameters may be very computationally expensive for large dimensional systems.

Once the macrostructure of the Artificial Neural Network has been initially framed, the inter- and intra-mechanics of the neuron nodes may be specified. Neuron operation is described by its input value operations and its function output. Input values to a node are determined by the connected weights and biases of the macrostructure (i.e., prior layers). These connection weights and biases are adjusted through the training of the Neural Network system. The best training algorithm will be based on its convergence speed, reduction of training error, and its performance on testing patterns. Adequate network training will also be a function of the number of epochs that the data is represented to the network for. Output operation is described by how the neuron fires vis-à-vis its activation function. Achieving best performance criteria for a given Artificial Neural Network architecture will depend on the level of success for optimization of weights, biases, and activation function(s). Testing will require adequate repetitions for each parameter value if initial connection weights are randomly seeded for each trial.

Section 4 presents results from performing these optimization trials as well as the results and recognition rate of the final trained classifier.

## 4 Testing Results

### 4.1 Neural Network Optimization

Initial construction of the Artificial Neural Network was based on typical specifications used for generic classifiers. A Multi-Layered Perceptron feed-forward network with 400 input nodes, 25 hidden nodes, and 3 output nodes was initiated. The MATLAB Neural Network toolbox was utilized to assist in the narrowing of appropriate Artificial Neural Network formats before creating customized code to provide a best solution for this particular problem. Training algorithms were executed for a maximum of 2000 epochs on 30 training images and subsequently tested on 24 test images (i.e., 8 astonished, 8 calm, and 8 smiling). Different categories of training algorithms were executed: conjugate gradient (b, cgb, cgf, cgp, and scg), quasi-Newtonian (oss), and standard backpropagation (gd, gda, gdm, gdx, and rp). The very own implementation of Multi-Layered Perceptron with BP network labelled "SYDE" has been included for convenient reference at this point. Fig. 2, Fig. 3, and Fig. 4 display the three diagonal elements of the confusion matrix. The correct levels for each of these figures are 8. In Fig. C1 in Appendix C - Statistical Analysis a full 3 by 3 confusion matrix box-plot is provided for further comparison of off-diagonal elements and statistical variation over the 12 repeats for each value.

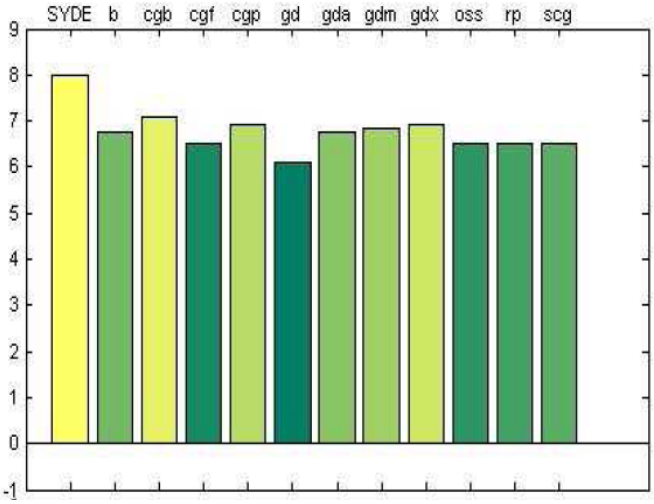



Fig. 2. Performance of Training Methods for Astonishment based on Logsig Activation, 25 Hidden Nodes, and 2000 Epochs (n=12)

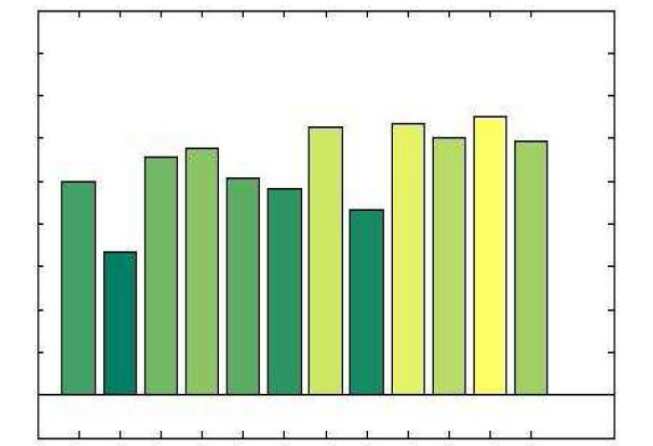


Fig. 3. Performance of Training Methods for Calm based on Logsig Activation, 25 Hidden Nodes, and 2000 Epochs (n=12)

The above training method plots reveal that the back propagation methods perform modestly better than the other available types. A modified back propagation training algorithm with acceleration, momentum, and a synthetically enlarged input dataset was implemented as the final "SYDE" method.

Next, optimization of the activation function was performed. Activation functions are categorized into two categories: differentiable, and non-differentiable. Since, back propagation methods rely on search lines formed from the error surface gradient, differentiable functions are preferred. The non-differentiable functions are: compet, hardlim, hardlims, and softmax. The differentiable functions are: logsig, netinv, poslin, purelin, radbas, satlin, and satlins. Fig. 5 through Fig. 7 display the analytical results.

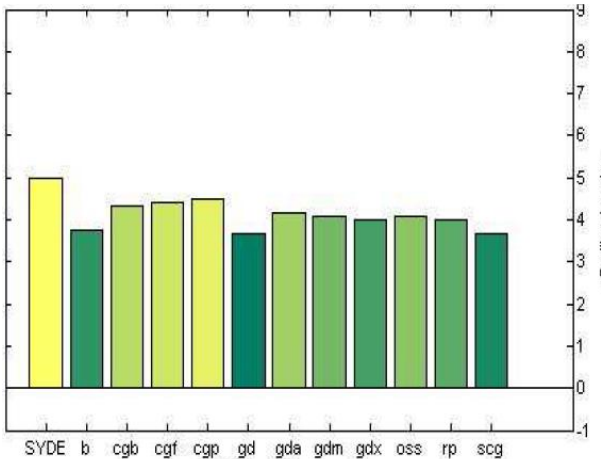



Fig. 4. Performance of Training Methods for Smiling based on Logsig Activation, 25 Hidden Nodes, and 2000 Epochs (n=12)

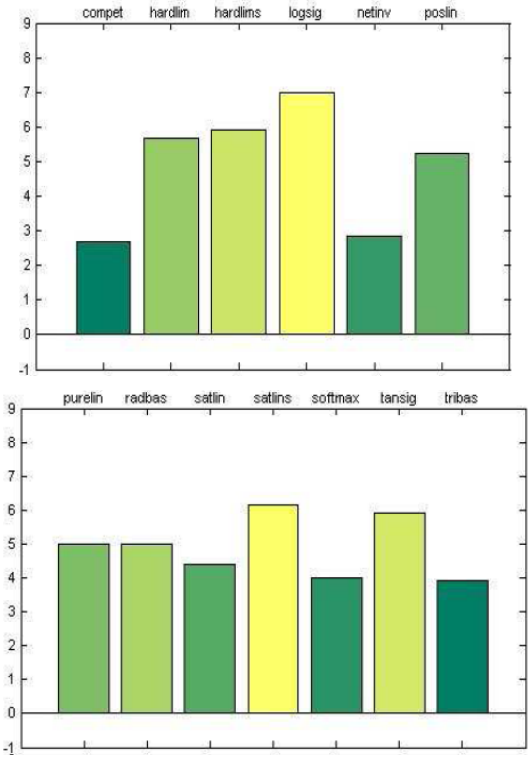



Fig. 5. Performance of Activation Functions for Astonishment based on Traingda Training, 25 Hidden Nodes, and 2000 Epochs (n=12)

In Fig. C2 and Fig. C3 in Appendix C - Statistical Results full 3 by 3 confusion matrix box-plots are provided for further comparison of off-diagonal elements and statistical variation over the 12 repeats for each value. As expected, the differentiable functions performed better than their counterparts. Both the logsig and tansig functions provided the best overall activation function performance values for the test set. The final implemented activation functions were the tansig function for the hidden layer and purelin for the output layer.

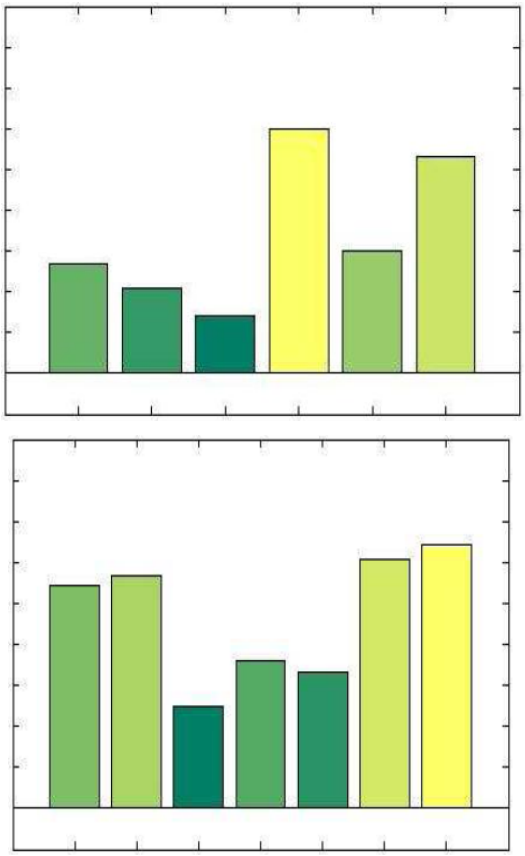


Fig. 6. Performance of Activation Functions for Calm based on Traingda Training, 25 Hidden Nodes, and 2000 Epochs (n=12)

The search testing performed provided a systematic method for determining an appropriate approach for a customized Neural Network. The customized training method began with a simple back propagation steepest gradient descent search line and then added more enhancements with the addition of accelerated learning rates and momentum terms. Activation scenarios centered on both the logsig and tansig activation functions. The tansig function provided a marginally better recognition rate. The purelin function for the output layer provided a means of mapping the hidden layer outputs directly (i.e., one-to-one) the output layer outputs. This composite Artificial Neural Network structure required a means of direct validation. Two common techniques of cross-validation are the K-fold and the one-left-out methods. For cross-validation all 54 images (previously 30 training and 24 test images) were pooled together and then divided into sets of 3. Over 18 trials (plus 5 repeats) subsequent sets of 3 were used as the test data and the remaining set of 51 used as the training set. The average for this cross-validation cycle was used as the benchmark for the performance index for this specification of the network parameters. This 18 by 5 cycle was performed 10 times for various hidden layer node number to determine an optimal number of hidden nodes. Fig. 8 exhibits that 25 hidden nodes in the single hidden node layer provide an optimal level of performance for this feed-forward Multi-Layered Perceptron

Artificial Neural Network architecture.

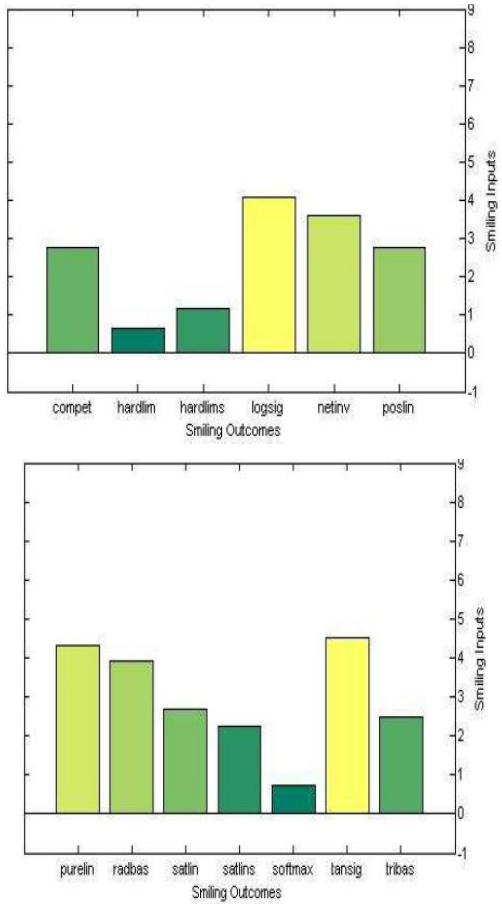



Fig. 7. Performance of Activation Functions for Smiling based on Traingda Training, 25 Hidden Nodes, and 2000 Epochs (n=12)

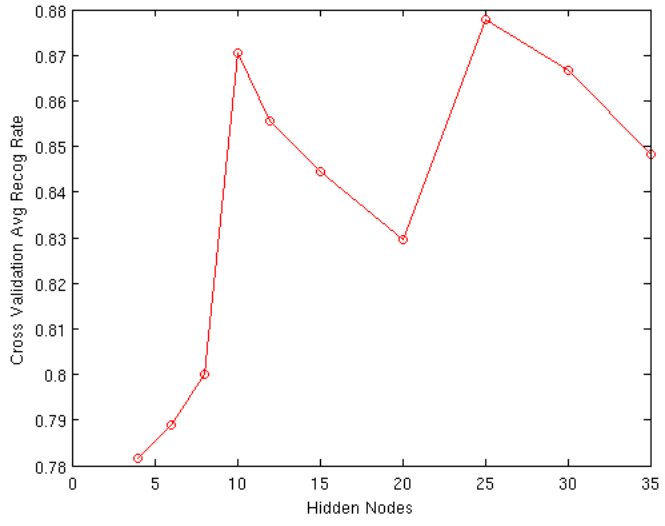



Fig. 8. Cross-Validation versus Hidden Node Number for Customized Training and Activation Functions, 2000 Epochs, Sets of 3 (n=5)

### 4.2 Evaluation Results

The final Neural Network classifier type is a 2 layer network with 25 hidden nodes. The activation functions were hyperbolic tangent functions or tansig. Training was performed using Gradient Descent with adaptation and momentum. The training epochs was set to 2000 to allow for convergence. The training set consists of 30 images total, 10 from each of the emotion categories. The test dataset contains 24 images, 8 from each of the emotion categories. The evaluation results consist of testing the trained network on unseen test images and tallying the percentage of error as well as the confusion matrix:

TABLE 1
CONFUSION MATRIX

| Confusion Matrix | Astonished (output) | Calm (output) | Smiling (output) |
|---|---|---|---|
| Astonished (true) | 8 | 0 | 0 |
| Calm (true) | 0 | 8 | 0 |
| Smiling (true) | 0 | 2 | 6 |

The recognition rate is the sum of the diagonals divided by the total sum and it is 91.7%.

## 5 CONCLUSIONS AND RECOMMENDATIONS

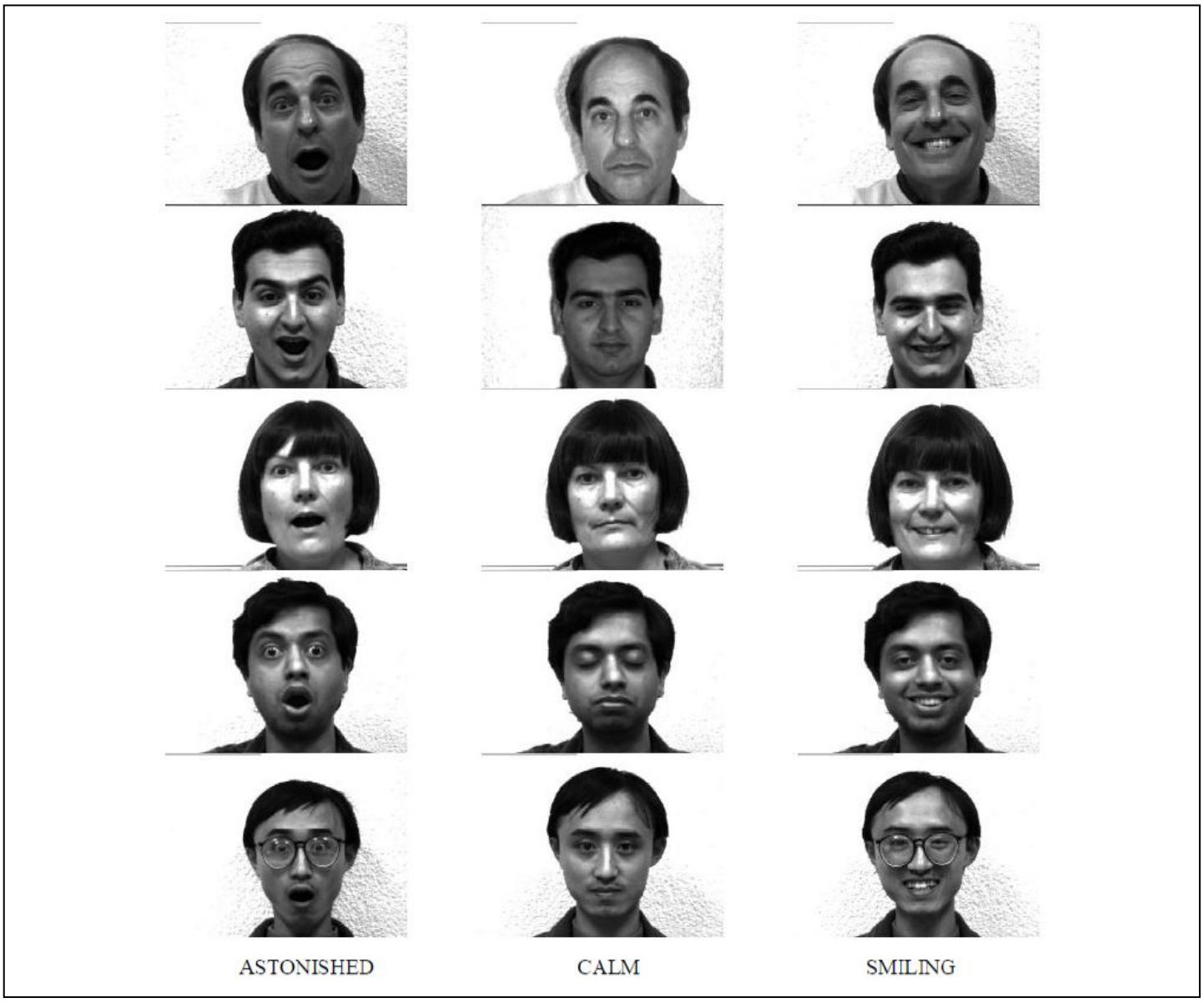
In this paper, emotion recognition using manually selected corners of a mouth is explored. The Neural Network classifier takes in image intensity and gives the emotion class in return. In the process of optimizing the parameters of the network, it is demonstrated via cross validation testing that the best number of hidden layers is around 25. Other optimum parameters for a network are the hyperbolic tangent activation function for the hidden nodes and the gradient descent + momentum + adaptation optimization method for learning.

It is also demonstrated that a possible solution to the problem of network over-fitting and an inability for the network to generalize/extrapolate well is to artificially create variants of the input mouth so as to enlarge the training data. This technique led to an increase in recognition rate. The overall result is very good as can be seen by the confusion matrix in section 4.2. The overall recognition rate for that test set is 91.7%

Further work in the area of expression recognition need to address the problem of using more information, specifically from the eyes. Also need to find a way to generate synthetic distorted facial images as a means to provide robustness for slight misalignment due to distortion of the facial muscles. Other extensions will include a way to automatically detect the key feature points to eliminate the manual position from the entire process.

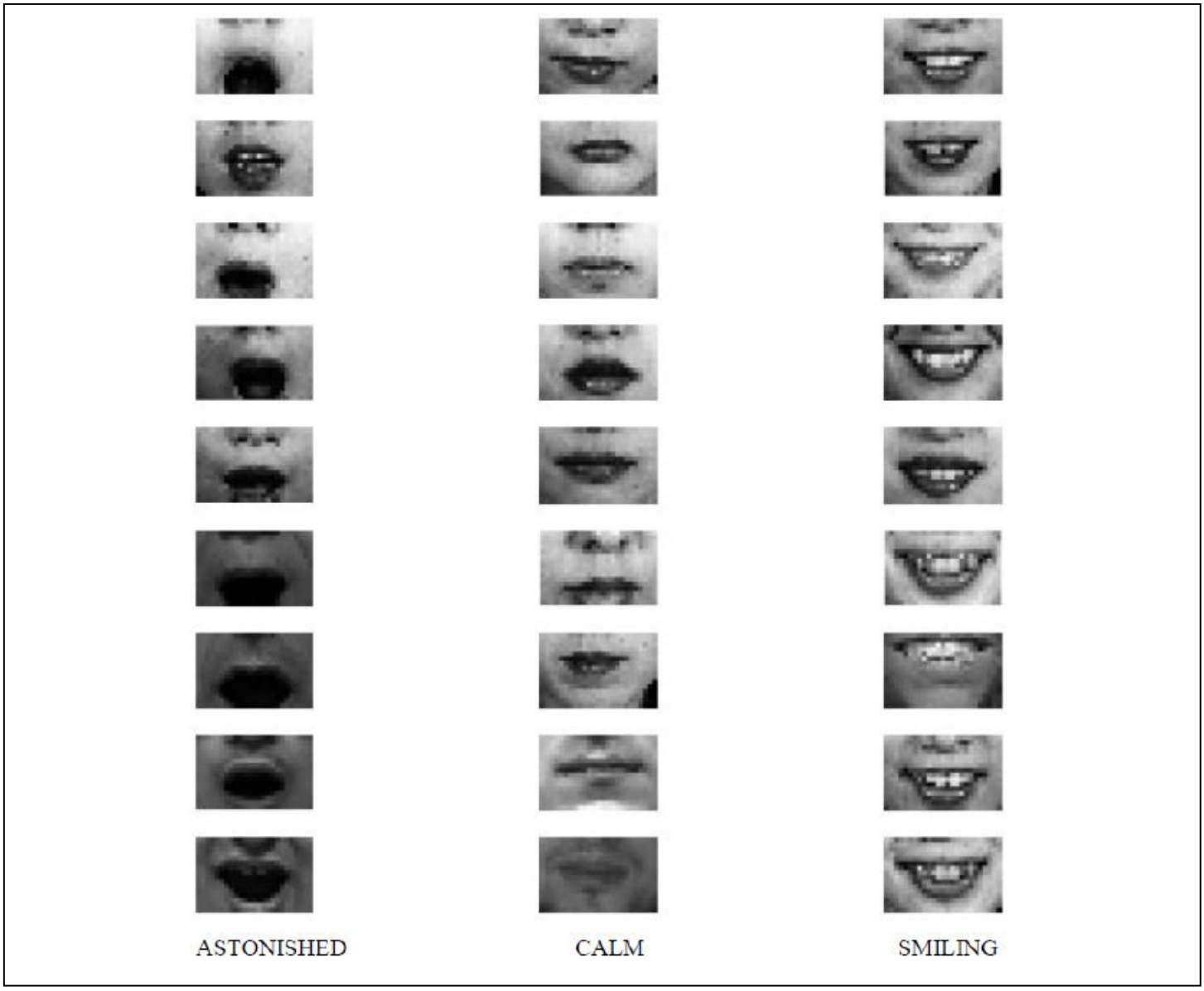
## APPENDIX A - RAW SOURCE IMAGES

The figures below are from the JAFFE and Yale Collections:



## APPENDIX B - PREPROCESSED INPUT DATA

The figures below are Input Data (400 Pixel) sorted by Target Clssification:



## APPENDIX C - STATISTICAL RESULTS

The figures below are Boxplots of Confusion Matrix Parameters:

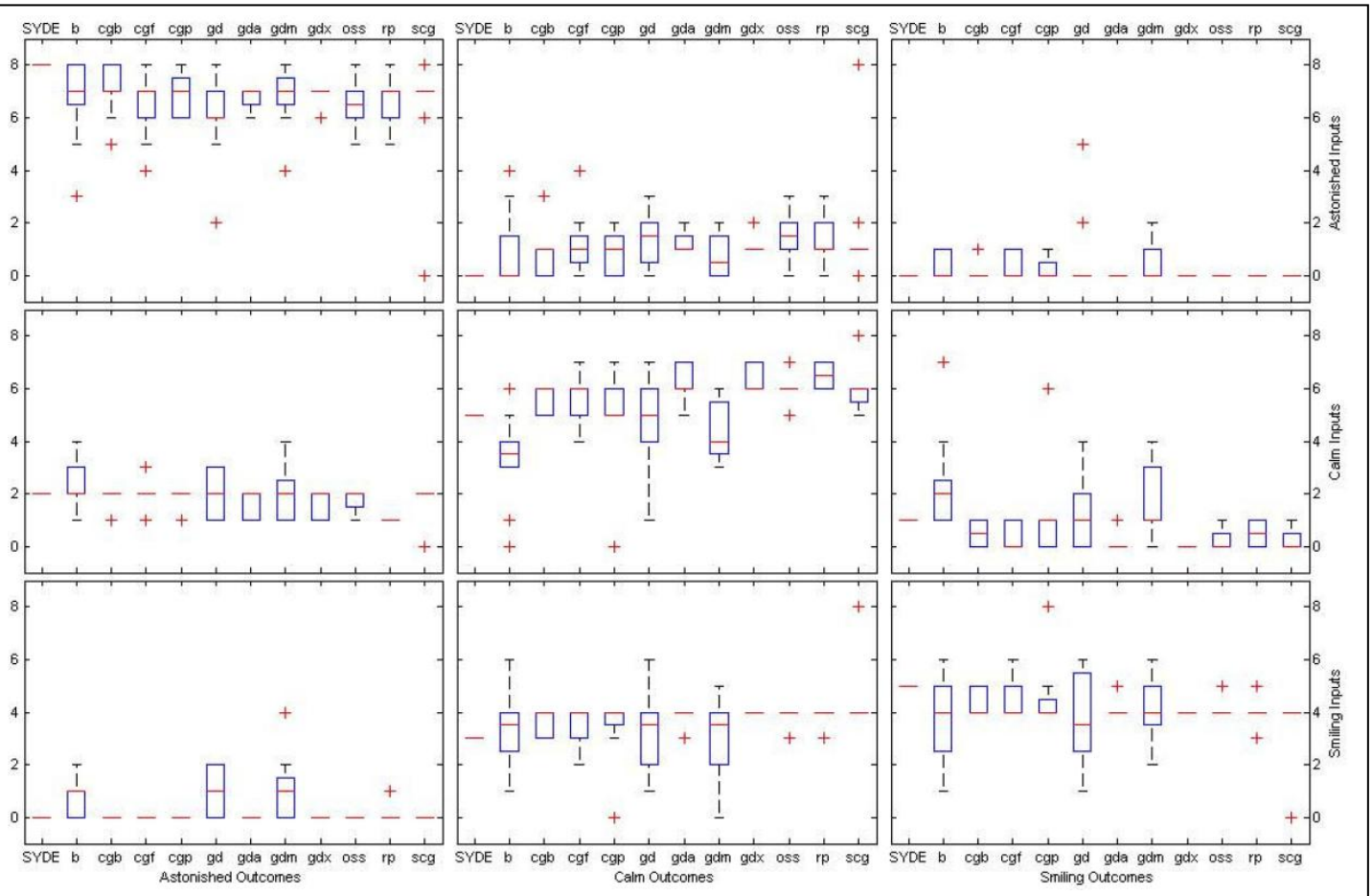



Fig. C1. Performance of Training Methods with logsig Activation, 2000 Epochs, 8 Astonished, 8 Calm, and 8 Smiling Inputs (n=12)

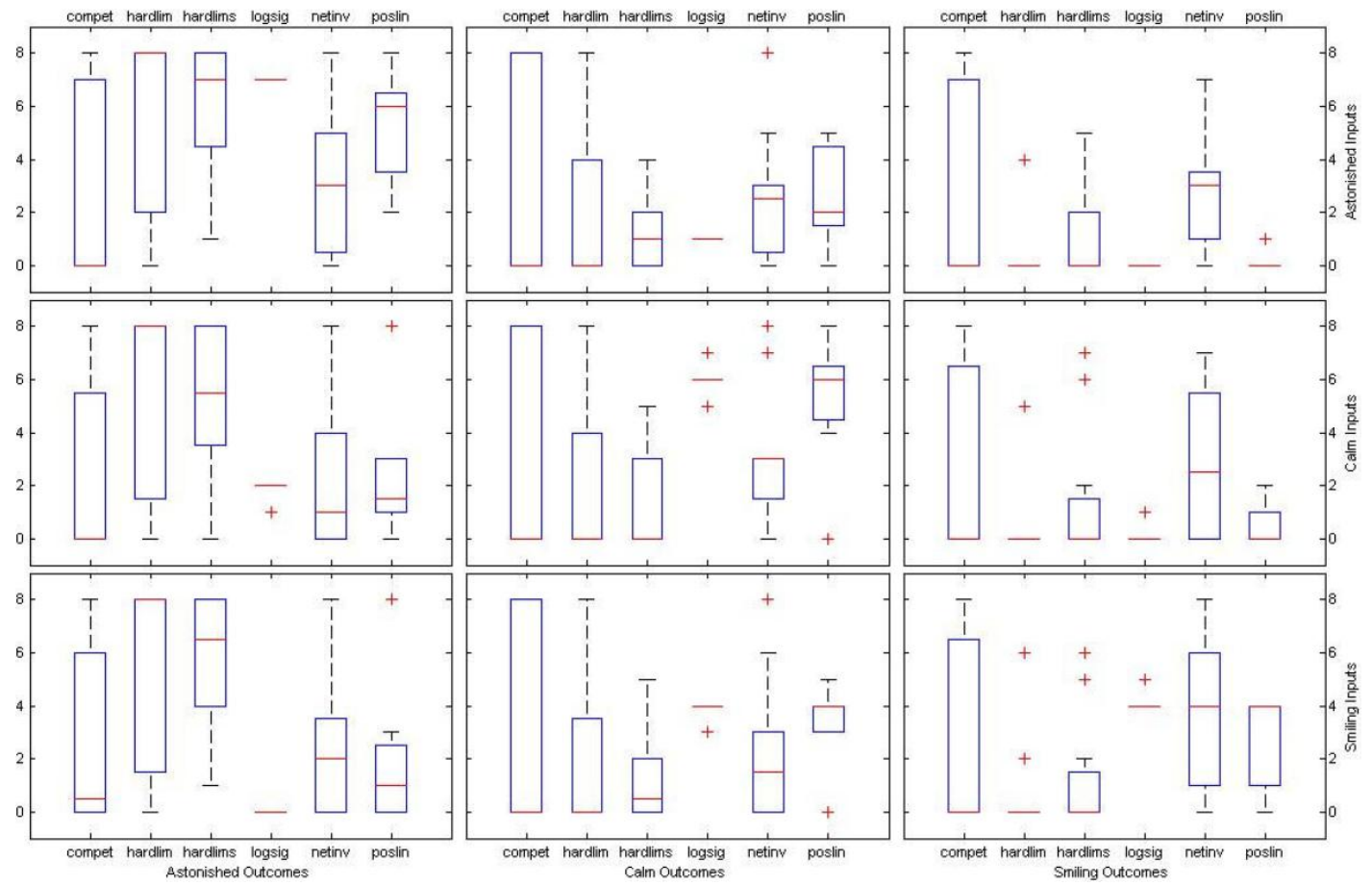



Fig. C2. Performance of Activation Functions (I) with Traingda Training, 2000 Epochs, 8 Astonished, 8 Calm, and 8 Smiling Inputs (n=12)

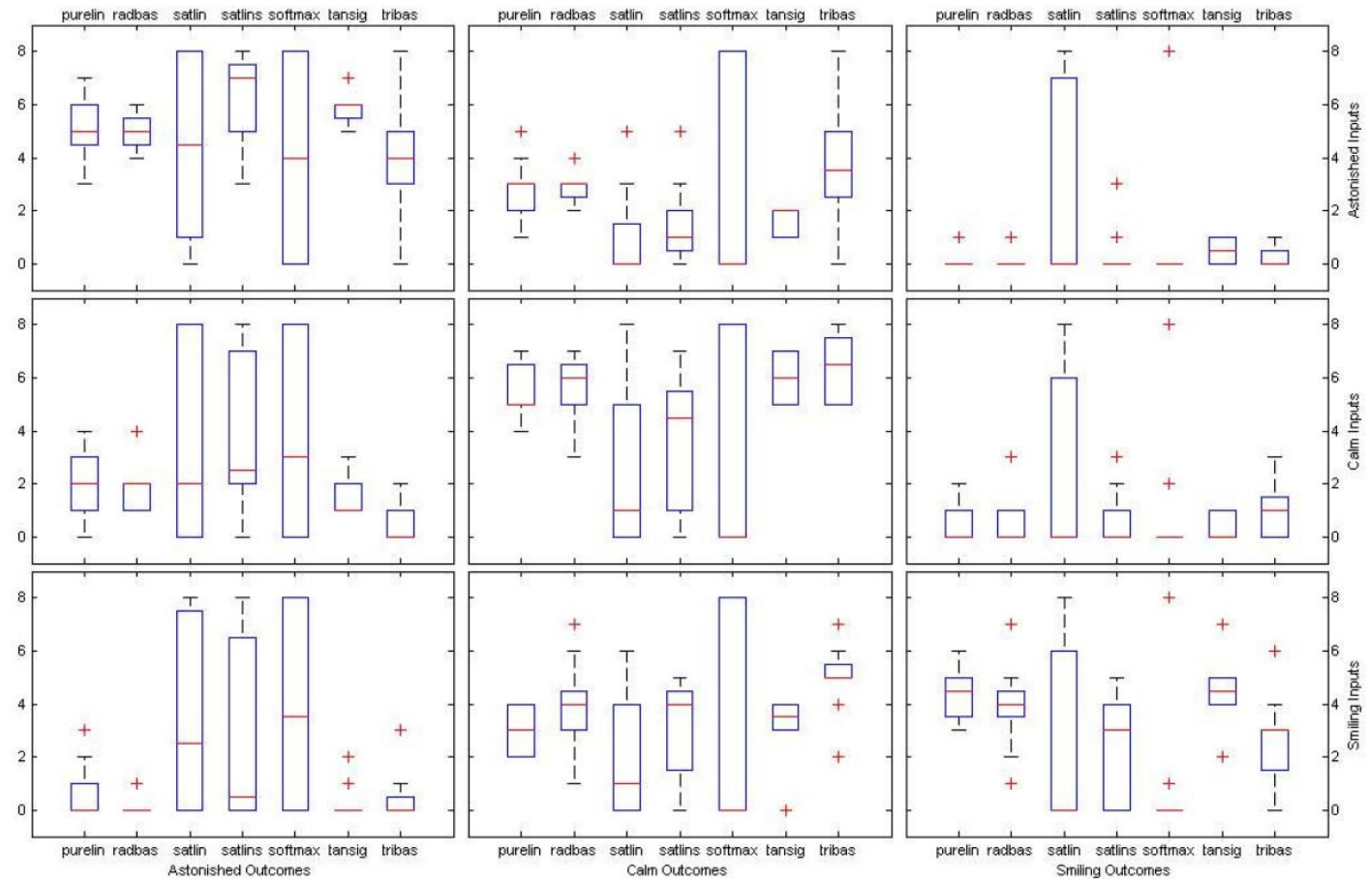



Fig. C3. Performance of Activation Functions (II) with Traingda Training, 2000 Epochs, 8 Astonished, 8 Calm, and 8 Smiling Inputs (n=12)

## REFERENCES

[1] Belhumeur, P N., and D J. Kriegman. "The Yale Face Database". 1997. Yale University, New Haven. 27 Mar. 2008 <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>.

[2] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi & Jiro Gyoba. "Coding Facial Expressions with Gabor Wavelets". Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition, April 14-16 1998, Nara Japan, IEEE Computer Society, pp. 200-205.

[3] Hancock, Peter. "Psychological Image Collection at Stirling (PICS)". University of Stirling, Stirling. 28 Mar. 2008 <http://pics.psych.stir.ac.uk/>.

[4] Lisetti and D. Rumelhart. "Facial Expression Recognition using a Neural Network". In Proceedings of the 11 th International Flairs Conference. AAAI Press, 1998.

[5] Jyh-Yeong Chang, and Jia-Lin Chen. "A facial expression recognition system using neural networks". IJCNN '99. International Joint Conference on Neural Networks, 1999.

[6] Franco, Leonardo, and Alessandro Treves. "A Neural Network Facial Expression Recognition System using Unsupervised Local Processing. "Image and Signal Processing and Analysis" (2001): 628-32. 31 Mar. 2008 <http://www.lcc.uma.es/~lfranco/B4-Franco+Treves01.pdf>

[7] Gu, L. and Kanade, T. 3D Alignment of Face in a Single Image, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, 2006.

[8] Zhou, Y., Gu, L. and Zhang, H. Bayesian Tangent Shape Model: Estimating Shape and Pose Parameters via Bayesian Inference, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Wisconsin, 2003.

[9] Bishop, Christopher. "Neural Networks for Pattern Recognition". London: Oxford University Press. 1995.

[10] Hopfield, J. (1984) "Neurons with graded response have collective computational properties like those of two state neurons," in *Proceedings of the National Academy of Science*, pp. 3088-92.