Gender Recognition of Human Behaviors Using Neural Ensembles

Jungwon Ryu and Sung-Bae Cho

Yonsei University
Department of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
rjungwon@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Abstract

In this paper, we have developed two ensembles of neural network classifiers in order to recognize actors' gender from their biological movements. One is the ensemble of modular MLPs (experts), the other is the ensemble of modular MLPs and an inductive decision tree which combines the output of experts. The human movement database consists of 13 males' and 13 females' movements, and contains 10 repetitions of knocking, waving and lifting movements both in neutral and angry style. Features have been extracted with 4 different representations such as the 2D and 3D velocities and positions, recorded from 6 point lights attached on body. We have compared the results of ensembles to the regular classifiers such as MLP, decision tree, self-organizing map and support vector machine. Furthermore, the discriminability and efficiency have been calculated for the comparison with the human performance that has been obtained with the same experiment. Our experimental results indicate that the ensemble models are superior to the conventional classifiers and human participants.

1 Introduction

The perception of biological motion has been an exciting field for psychological researchers over the last three decades. Especially, the perception of the human behavior also has been the interesting topic for the computer scientists. In order to design systems that act human-like, such as humanoid robot or avatar, we need the interdisciplinary research on the human and computer interaction, which has been highly spotlighted these days [1]. In this paper, we attempt to propose the optimal model of machine learning classifiers by examining the ability of recognition of human's gender.

We propose two ensemble models to classify actor's gender from his arm movements, such as knocking, waving and lifting done in both neutral and angry style. One is the ensemble of modular multilayer perceptrons, and the other is the ensemble of decision tree and modular multilayer perceptrons. The benefit of the ensemble of classifiers is that the combination of several prospective models may produce better prediction [2]. Since choosing a single best method for a given problem results in wasting models that lost the competition, we can also take advantages of utilizing the networks constructed under the modularity principle with respect to the granularity of information of data.

We have used 4 representations (features) of the input patterns, such as velocities and positions in 2 and 3 dimensional spaces. We have compared the results of ensembles with those of conventional classifiers and human results by obtaining the discriminability and efficiency [3].

2 Background

2.1 Gender Recognition

W. Wolff first studied how people can recognize friends by their walk [4]. In order to overcome the confounding role of familiarity cues such as size of shape of objects, G. Johanssen approached this problem with the moving spots and lines [5, 6]. In his experiment, he used glass-bead retro reflective tape attached on the main joint of the human body at 10 different sites. Filmed displays of the point light walkers that only the lights reflected from the tape can be seen as illuminated dots in a dark background were used to test people's perception of biological motion. G. Johanssen could not extract any helpful information for the recognition from the static configuration of point light displays, but once they are moving, people could recognize the walker somehow. Figure 1 is an example of point light display of people drinking.

In Cutting and Kozlowski's research the sex of the walkers could be recognized from displays of point light sources mounted on people's major joints with an average of 60~70% accuracy [7]. They also found that lights on the

upper body's joints are more useful to recognize the movement.

In this context, we concentrate on the moving spots on the upper body's movement, such as 'knocking,' 'waving' and 'lifting', from the dynamic point light display as the classification cue. These movements are good experimental objects because they have the short time duration, easy to be played so that take place often to be experienced easily.

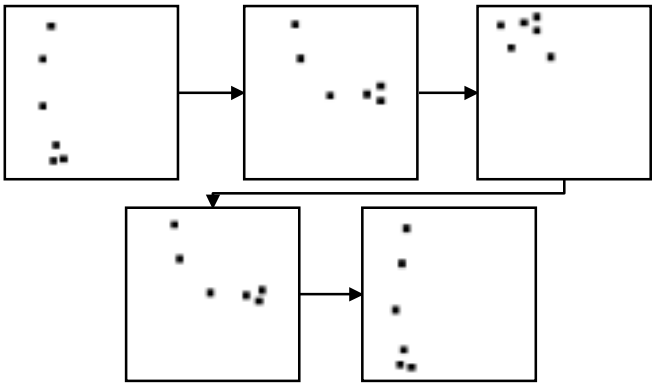

Figure 1: Point light display of people drinking

2.2 Mixture of Expert

According to the Osherson's definition, a neural network is said to be modular if the computation performed by the network can be decomposed into two or more subsystems that operate on distinct inputs without communicating with each other [8]. After this notion of modular connectionist systems was first discussed in the mid of 1980's by Barto and Hinton, Pollack proposed the cascaded backpropagation architecture [9] and Jacobs developed taxonomy for a class of modular hierarchical connectionist models [10]. Hamshire and Waibel have proposed the Meta-Pi, which consists of a number of source-dependent sub networks that are integrated by a combinational time-delay neural network [11], Lincoln and Skrzypek proposed clustering multiple backpropagation networks [12], Battiti and Colla suggested the concept of democracy to combine the outputs of different neural network classifiers [13]. These early examples have shown that integrating the multiple modules, often referred as committee machines, could have enhanced the accuracy and generalization capacity.

More recently, Gutta *et al.* suggested the hybrid models of RBF networks and decision trees with FERET face image database for gender, ethnic origin and pose of face classification [2]. They have put Gaussian noise to the original image and done 5 degree of geometric transformation as the input of hybrid classifier and obtained 93.3 percent correct of the gender classification over the 60 test sets. In the current research, we have examined the recognition ability of neural ensembles with 4 features of same movements, not just geometrical transformation but such as 2D, 3D velocities and positions.

3 Human Movement Data

3.1 Data Acquisition

The movement data were obtained using a 3D position analysis system (Optotrak, Northern Digital). Positions of the head, right shoulder, elbow, wrist and the first and fourth metacarpal joints were recorded at a rate of 60 Hz while an actor performed the movement. Actors were instructed to perform knocking, waving, and lifting movements in neutral and angry style. 26 people participated to build the data set, half of them were males and the others were females. For each of the 6 combinations of motion and affect was recorded 10 times repeatedly, 8 of them have been used as the training data and the rest as the test data.

3.2 Preprocessing

Each movement has been processed to obtain the start and end point. The start of movement was defined as the instant the tangential velocity of the wrist rose above the 5% of the peak and the end by the instant after the velocity passed below the 5% of the peak. The missing data that had been resulted from a marker going out of view of the cameras have been interpolated to remove this artifact.

At this point, since we only have the 3D position data of movements, the transformation of data representation has been done. We have used 4 representations of movement, which denoted as 2Dvel, 2Dpos, 3Dvel and 3Dpos. The 2Dvel is the velocity on the *y*-*z* plane and the 2Dpos is the position on the *y*-*z* plane. The 3Dvel and 3Dpos are the velocity and position in 3-dimensional (*x*, *y*, *z*) space.

It is also necessary to normalize the patterns so that the data stay on the standard levels of length and amplitude. In this paper, we have normalized all the patterns to 150. As can be seen in Figure 2, even after the normalization, the characteristic of pattern still has been preserved well.

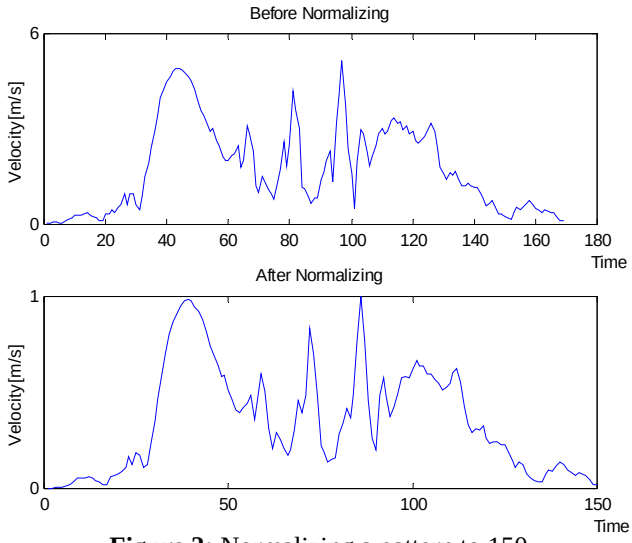

Figure 2: Normalizing a pattern to 150

An example of the final set of input patterns has been shown in Figure 3. This is the 2Dpos training set and contains 1,248 patterns, each of which has 1,800 dimensions. We can see there are the 3 different types of patterns, the first 416 patterns were knocking, the next were waving and the last were lifting movements. There are also the neutral and angry styles mixed, but it is not clearly distinguishable.

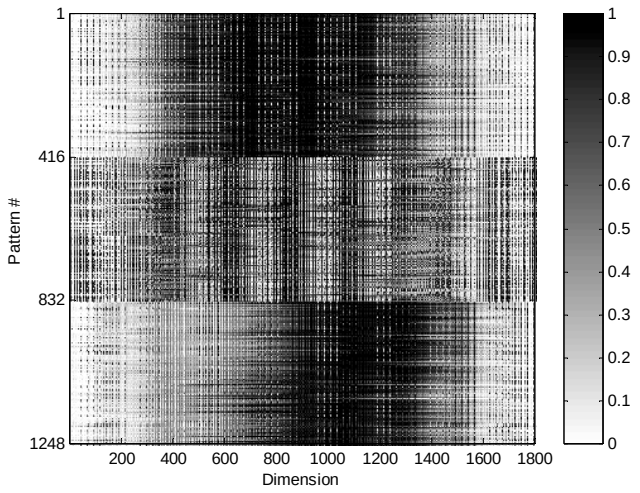

Figure 3: Learning data set (2Dpos)

4 Neural Ensembles

4.1 Ensemble of Modular MLPs

Multilayer perceptron (MLP) is commonly used in such field of pattern recognition due to its powerful and stable learning algorithms [14]. The backpropagation based on the delta learning algorithm is a good example [15, 16, 17]. The computing power of the backpropagation algorithm lies in two main attributes: local method for updating the synaptic weights and biases of the multilayer perceptron and efficient method for computing all the partial derivatives of the cost function with respect to these free parameters [17].

In the 3-layered MLP, given input vector X, the hidden neuron (Z) and output neuron (Y) are activated by following equations:

$$Z = f[Net(w_1 X)], \quad Y = f[Net(w_2 Z)]$$

where w_1 and w_2 are the weights between first and second, second and third layer, $Net(.)$ is weighted sum of nodes connected from the prior layer and $f[.]$ is a sigmoid activation function. Then the errors on the output and hidden layer are expressed as:

$$\delta_2 = (d - Y) f'[Net(w_2 Z)], \quad \delta_1 = \left(\sum \delta_2 w_2\right) f'[Net(w_1 X)]$$

where d is the desired output. Then weights are updated by:

$$w_2(t+1) = w_2(t) + \alpha \delta_2 Z, \quad w_1(t+1) = w_1(t) + \alpha \delta_1 X$$

α is a learning constant. These procedures are iterated until the network reaches on a certain level of recognition.

The ensemble of modular MLPs (EMMLP) proposed in this paper is shown in Figure 4. Since there are six combinations of affect and motion conditions, we have divided the whole data into six sub classes to reflect the locality of our data. The EMMLP consists of an affect-motion classifier has been trained to recognize the corresponding sub class, and six modular MLPs which are in charge of each of sub classes. Once the given input pattern is assigned to a certain class, the modular MLP of that class makes the decision. All modular networks have been trained separately only with the data that they are in charge of.

Of course, EMMLP can perform best if the affect and motion classifier gives perfect answer, but it takes advantages that even misclassified patterns by the affect-motion classifier could be correctly classified by chance.

Each MLPs have 900~2700 input, 50 hidden and 2 output nodes. The learning rate and momentum have been chosen experimentally. All the networks have been trained until the recognition rate on the training data reaches on 98.0%. “Male” is encoded as [1 0] and “female” as [0 1] on the output layer.

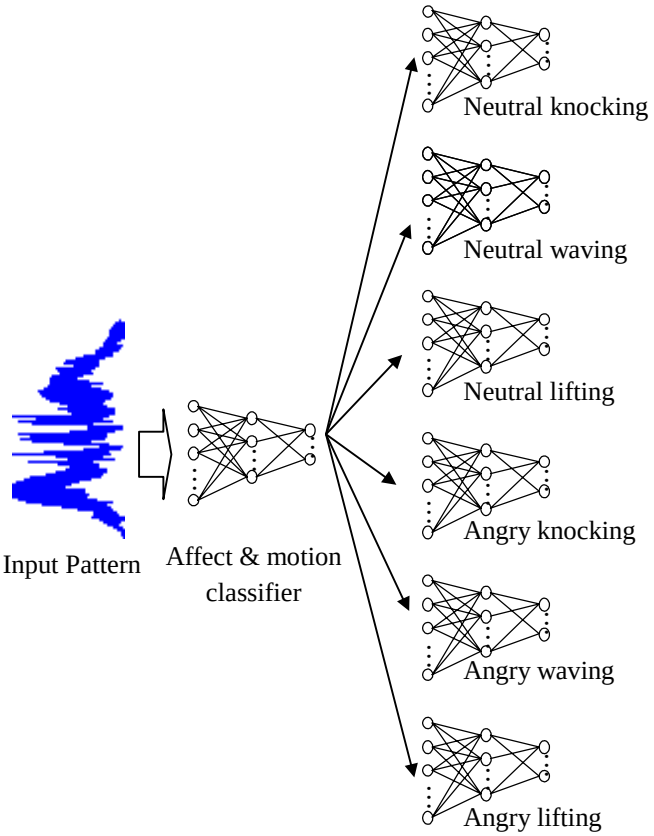

Figure 4: Ensemble of Modular MLPs

4.2 Ensemble of Modular MLPs and Decision Tree

The basic aim of concept-learning induction system, such as decision tree (DT) is to construct rules for the classification

from the set of objects of which class labels are known [18].

Quinlan's C4.5 uses an information-theoretical approach based on the energy entropy. C4.5 builds the decision tree using a divide-and-conquer approach: select an attribute, divide the training set into subsets characterized by the possible values of the attribute, and follow the same partitioning procedure recursively with each subset until no subset contains objects from more than one class. The single class subsets correspond them to the leaves. The entropy-based criterion that has been used for the selection of the attribute is called the gain ratio criterion [19].

Let X be a possible test (attribute selection) that partitions the training set T into n sub sets (T_1, T_2, …, T_n), split_info(X) as the entropy of a message where information is given in terms of outcomes, and gain_ratio(X) as:

$$\text{split\_info}(X,T) = -\sum\left(\frac{|T_i|}{|T|}\right)\log_2\left(\frac{|T_i|}{|T|}\right)$$

$$\text{gain\_ratio}(X) = \frac{\text{gain}(X)}{\text{split\_info}(X)}$$

The gain ratio criterion selects that test X such that the gain ratio(X) is maximized.

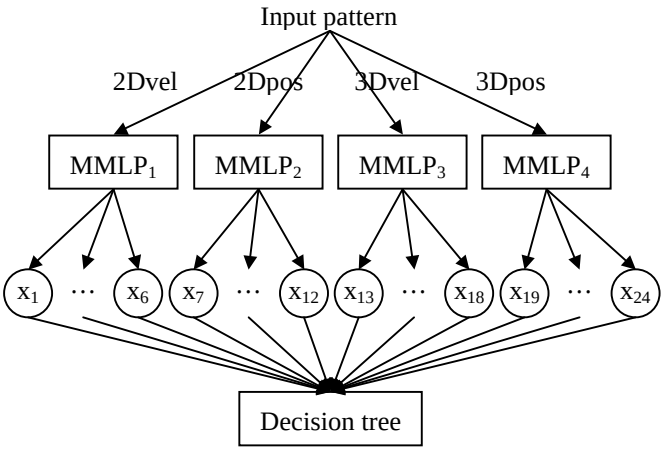

Figure 5: Ensemble of modular MLPs and decision tree

The second classifier, ensemble of modular MLPs and decision tree (EMMLP/DT) is as shown in Figure 4. We have used 5 data representations at the same time to discriminate the actor's gender. $MMLP_1$ consists of 6 modular networks trained only with 2D velocity data, $MMLP_2$ with 2D position, $MMLP_3$ with 3D velocity and $MMLP_4$ with 3D position data in the same way as those of EMMLPs. Given the input pattern, the 4 MMLPs have their 6 outcomes per each, and finally the output vector (x_1, x_2, …, x_{24}) has been chosen for training, which is the input to the decision tree.

The decision tree has decision nodes and leaves. Leaves indicate the class label such as "male" or "female" and the decision nodes specify some test to be carried out on a single attribute value, with one branch for each possible outcome of the test [20]. A people's movement is an object, the attributes are the dimensions of input vectors, that are continuous numbers normalized between 1 and 0.

The benefits of using decision tree to combine the outcomes of MMLPs can be summarized as the use of the multiple modalities of human movement, the flexibility and adaptivity of thresholds derived using the entropy as opposed to *ad hoc* and hard thresholds and the intuitional interpretability of the result of classification.

5 Other Classifiers

5.1 Support Vector Machine

The Support vector machine (SVM) introduced by V. Vapnic in 1995 is a method to estimate the function classifying the data into two classes [21]. The basic idea of SVM is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized. SVM achieves this by the structural risk minimization principle that is based on the fact that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension [18].

Given a labeled set of M training samples (X_i, y_i), where $X_i \in R^N$ and y_i is the associated label, $y_i \in \{-1, 1\}$, the discriminant hyperplane is defined by:

$$f(X) = \sum_{i=1}^{M} y_i \alpha_i k(X, X_i) + b$$

where $k(.,.)$ is a kernel function and the sign of $f(X)$ determines the membership of X. Constructing an optimal hyperplane is equivalent to finding all the nonzero [22]. This paper has used SVM^{light} module that imports quadratic programming techniques.

5.2 Self-organizing Map

The Self-organizing map (SOM), by T. Kohonen, defines a mapping from the input data space onto an output layer by using Kohonen's unsupervised learning algorithm [23, 24].

The SOM has an input layer and output layer. The output layer consists of N nodes, each of nodes represents a vector that has the same dimensions as the input patterns. For given input vector X. the winner node m_c is chosen using the Euclidean distance between X and neighbor nodes m_i. Then, updating the winner node's weight vector takes place.

$$\|x - m_c\| = \min_i \|x - m_i\|$$

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\}$$

where, $\alpha(t)$ is the learning rate, $n_{ci}(t)$ is a neighborhood function. We have used 10 by 10 map, rectangular topology, and 0.02 of the learning rate.

5 Results

5.1 Recognition Rates

The final result of recognition rates with respect to the classifiers and the data representations has shown Table 1. Experiments with ensembles have been repeated 3 times.

Table 1: Recognition rate of classifiers [%]

Classifier	Data Representation				Aver-age
	2Dvel	2Dpos	3Dvel	3Dpos	
EMMLP	80.0	84.6	80.8	86.5	82.9
EMMLP/DT	81.4				81.4
MLP	75.2	59.6	81.4	84.6	75.2
SVM	68.6	75.0	71.8	73.2	72.2
SOM	65.7	76.6	60.6	76.9	70.0
DT	70.2	69.2	67.0	72.8	69.8
Human	51.3				51.3

On the average, the EMMLP performed best and the EMMLP/DT was the second, followed by MLP, SVM, SOM, DT, and human participants in order. However, even for one classifier, the performance varies on the used representation. The recognition rate of EMMLP varies from 80.0% to 86.5%. Since the EMMLP/DT uses all representations at the same time, it has just one number.

Among the data representations, most classifiers, except SVM, have obtained the best recognition rates when 3Dpos was used. This indicates that 3Dpos is the most informative data representation to classify gender.

Figure 6 shows a part of the decision tree constructed by EMMLP/DT method. The left branches of each decision nodes are the case when the attribute value is larger than the criterion number centered between branches and the right branches are when the attribute value is smaller. A pattern that has larger x_{18} than 0.4812 and smaller x_{17} than 0.0027, it will be classified as a female's movement. The whole tree size was 35.

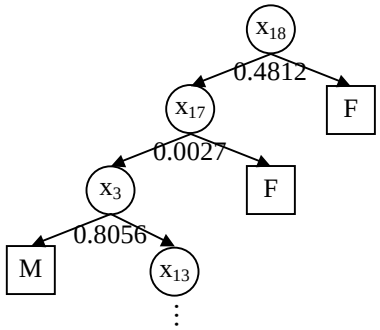

Figure 6: Result of EMMLP/DT

Generally, machine learning classifiers performed much better than human. While the EMMLP yields up to the 86.5% of recognition rate, human's was just 51.3%.

5.2 Discriminability and Efficiency

In the signal detection theory, the discriminability (or sensitivity) and efficiency are often used when measuring the receiver's capacity of signal discrimination. Suppose an observer is forced to indicate whether or not the light was flashed, there can be four possible cases with respect to his response on the given input signal.

Table 2: Four possible cases

		Input Signal	
		Signal	Noise
Response	Yes	Hit	False alarm
	No	Miss	Correct rejection

Based on the numbers got from the confusion matrix above, we can draw the internal response probability distribution curves of his observation (Figure 7). The discriminability can be thought of the distance between two peaks. The high d' indicates the classifier performs well.

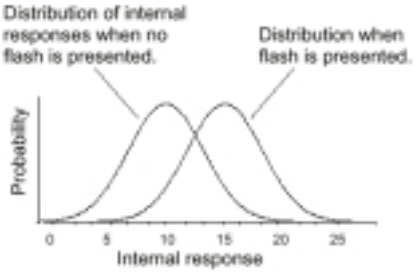

Figure 7: Internal response probability distributions

If the observer 1's discriminability and observer 2's are d'_1, d'_2, respectively, the efficiency of observer 1 over observer 2 is defined as [25]:

$$E_{12} = \left(\frac{d'_1}{d'_2} \right)^2$$

For our experiments, we have assumed that 'signal' is when the pattern is male's, and 'noise' is when female's.

Table 3: Discriminability of classifiers

Classifier	Data Representation				Aver-age
	2Dvel	2Dpos	3Dvel	3Dpos	
EMMLP	1.20	1.30	1.07	1.22	1.20
EMMLP/DT	1.03				1.03
MLP	0.73	0.27	0.98	0.75	0.68
SVM	0.62	0.77	0.70	0.66	0.69
SOM	0.54	0.91	0.37	1.01	0.71
DT	0.71	0.66	0.62	0.73	0.68
Human	0.13				0.13

Table 3 is the result of discriminability. According to the average, the EMMLP gave the best result, and the EMMLP/DT was the second. Similarly to the recognition

rate results, the 3Dpos was the most distinguishable representation. EMMLP's 2Dpos have obtained the highest value, 1.30.

Table 4: Efficiency of ensemble classifiers

Classifier	Data Representation				Aver-age
	2Dvel	2Dpos	3Dvel	3Dpos	
EMMLP	85.2	100.0	67.7	88.1	85.3
EMMLP/DT	62.8				62.8

The efficiency of ensemble classifiers against the human participants is as shown in table 4. EMMLP with 2Dpos obtained highest efficiency of 100.0. It is simply because EMMLP with 2Dpos has the highest d'.

6 Concluding Remarks

In this paper, to classify the human's gender, we have proposed two ensemble classifiers, named as EMMLP and EMMLP/DT and compared the performance with other conventional classifiers. As the result, the EMMLP performed best and the EMMLP/DT was the second both in recognition rate and discriminability. Even though the simple MLP and DT obtained approximately 75% and 70% of recognition rate, we have enhanced the performance up to 86.5% combining modular MLPs and ensemble of the modular MLPs and decision tree. Among representations, 3Dpos was most informative (best feature), most of classifiers have performed with this type of representation. The results of efficiency have shown the neural ensembles are much good at recognizing the actor's gender than human participants.

References

[1] J. Ryu and S.B. Cho, "Searching for optimal features for gender recognition with neural network classifier," *Proc. of Intl. Conf. on Soft Computing*, Iizuka, October, 2000.

[2] S. Gutta, R.J. Huang, P. Jonathon and H. Wechsler, "Mixture of experts for classification for gender, ethnic origin, and pose of human faces," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, July, 2000.

[3] F.E. Pollick, V. Lestou, J. Ryu and S.B. Cho, "Estimating Efficiency in the Categorization of Biological Motion," *Vision ScienceS Society*, May, 2001.

[4] W. Wolff, *The Expression of Personality: Experimental Depth Psychology*, New York, Harper & Brothers, 1943.

[5] G. Johanssen, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, pp. 201-211, 1973.

[6] G. Johanssen, "Visual motion perception," *Scientific American*, vol. 232, pp. 76-89, 1975.

[7] J.E. Cutting and L.T. Kozlowski, "Recognising friends by their walk: Gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353-356, 1977.

[8] D.N. Osherson, S. Weinstein and M. Stoli, "Modular learning," *Computational Neuroscience*, E.L. Schwartz, Ed., pp. 369-377, Cambridge, MA: MIT Press, 1990.

[9] J. Pollack, "Cascaded back-propagation on dynamic connectionist networks," *Proc. Ninth Ann. Conf. Cognitive Sci., Soc.*, pp. 391-404, 1987.

[10] R. Jacobs, "Initial experiments on constructing domains of expertise and hierarchies in connectionist systems," *Proc. 1988 Connectionist Models Summer School* (San Mateo, CA), pp. 144-153, 1988.

[11] J.B. Hamshire and A. Waibel, "The Meta-Pi network: Building distributed knowledge representations for robust multisource pattern recognition," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 14, pp. 751-769, July, 1992.

[12] W.P. Lincoln and J. Skrzypek, "Synergy of clustering multiple backpropagation networks," *Advances in Neural Information Processing Systems*, D.S. Touretzky, Ed. San Mateo, CA: Morgan Kauffmann, vol. 1, pp.650-657, 1990.

[13] R. Battiti and A.M. Colla, "Democracy in neural nets: voting schemes for classification," *Neural Networks*, vol. 7, pp. 691-707, July, 1994.

[14] R.P. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4-22, April, 1987.

[15] B.V.S. Murthy, "Delta learning law for a single neuron," *IEEE Intl. Joint Conf. on Neural Networks*, vol. 3, pp. 1779-1782, 1999.

[16] P.J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," *The Roots of Backpropagation*, NewYork, John Wiley & Sons, Chapter 11, 1994.

[17] H.D. Beale, *Neural Network Design*, PWS Publish Company, pp. 11:1-47, 1996.

[18] J.R. Quinlan, "The effect of noise on concept learning," Machine Learning: an Artificial Intelligence Approach, R. S. Michalski, J. G. Carbonell and T. M. Mitchell, Des. San Mateo, CA: Morgan Kauffmann, vol.2, pp. 149-166, 1986.

[19] A.J.C. Sharkey, "On combining neural nets," *Connection Science*, vol. 8, pp. 299-313, 1996.

[20] S. Haykin, *Neural Networks*, pp. 318-350, 2nd Ed., Prentice Hall, 1999.

[21] V.N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer, 1995.

[22] B. Moghaddam and M.H. Yang, "Gender classification with support vector machines," *Proc. of 4th IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2000.

[23] T. Kohonen, "The self-organizing map," *Proc. of IEEE*, vol. 78, no. 9, pp. 1464-1480, September,1990.

[24] T. Kohonen, *Self-organizing Maps*, Springer, Berlin Heidelberg, 1995.

[25] H.B. Barlow, "The efficiency of detecting changes in random dot patterns," *Vision Research*, vol. 18, pp. 637-650, 1978.