GENDER RECOGNITION OF HUMAN BEHAVIORS USING ENSEMBLES OF NEURAL NETWORK CLASSIFIERS[1]

J Ryu, S-B Cho

Department of Computer Science, Yonsei University, Korea
rjungwon@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Received: 15/09/2001
Accepted in the final form: 06/02/2002

Abstract. In this paper, we propose two neural network classifiers that are composed of local experts to solve the problem of human gender recognition. Two ensembles of neural network classifiers have been proposed to discriminate the human gender by using the information of moving joints of actors. One is a hierarchical ensemble of modular MLPs (experts) and the other is the ensemble of modular MLPs by an inductive decision tree that combines the output of experts. The database consists of 13 males' and 13 females' human movements and contains 10 repetitions of knocking, waving and lifting movements both in neutral and angry styles. Features have been extracted from 4 data 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 with the conventional classifiers such as single MLP, decision tree, self-organizing map and support vector machine. Furthermore, the discriminability and efficiency ratio have been calculated for the comparison with the human performance that has been obtained from the same data. The experimental results indicate that the ensemble classifiers are much superior to the conventional classifiers and human subjects.

Keywords: Gender Recognition, Human Movement Recognition, Neural Networks, Mixture of Experts, Point Light Display

1 Introduction

The perception of the human behavior has been an interesting research topic for the computer scientists. It is because we need an interdisciplinary research on the human behavior, in order to design more intelligent systems that act and interact in a human-like manner. For this reason, the so-called 'human and computer interaction (HCI)' has been highly spotlighted these days.

The perception of biological motion has also been an exciting field for psychologists over the last three decades. Wolff first studied how people can recognize friends by their walk [1]. In order to overcome the confounding role of familiarity cues such as size and shape of objects that have been considered as important from his work, Johanssen approached this problem with the moving spots and lines [2, 3]. In his experiment, he used glass-bead retro reflective tape attached on the main joint of the human body at 10 different points. 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. Johanssen could not extract any helpful information for the

[1]This paper was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Science and Technology.

recognition from the static configuration of point light displays, but once they are moving, people could recognize the walker somehow. Much research, such as tracking and object seperating, has been proceeded since his work using moving point lights analysis [4, 5, 6]. Figure 1 is an example of point light display of people drinking a cup of water on the table. Additionally, 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 observed that lights on the upper body's joints were more useful to recognize the movement.

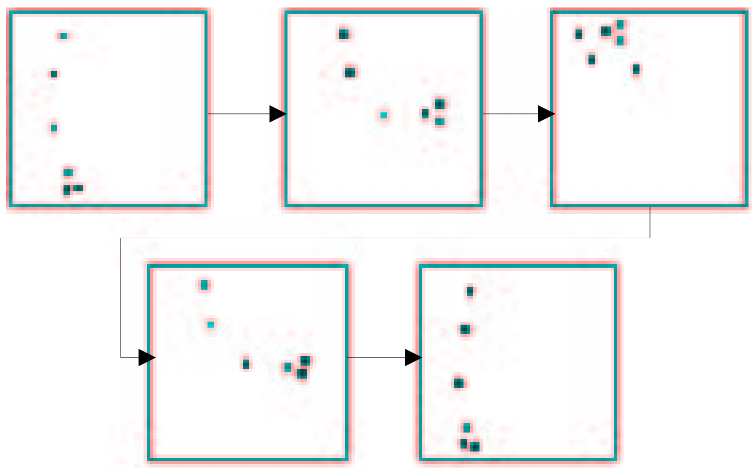

Figure 1: Example of point light display

In the field of computer science, many promising machine learning techniques and algorithms have been successfully adopted in many pattern classification and recognition problems [8, 9]. However, there still exist many difficulties to deal with real world problems. Some cases have too small number of samples and some cases have the data sets biased to certain classes, which makes the problem very difficult to solve. In those cases, the ensembles of multiple classifiers are often used. It may not always be a perfect solution for all the real world problems, but combining set of classifiers, like committee machines, provides more robust performance of classification [10].

In this paper, we propose the optimal model of neural ensemble classifiers by examining the ability of recognizing human's gender. We have two models to classify actor's characteristics through his arm movements, such as knocking, waving and lifting done in both neutral and angry style. One is a hierarchical ensemble of modular multilayer perceptrons, with the other is the ensemble of modular multilayer perceptrons and decision tree. The major rationale of the ensemble classifiers lies in the better prediction [11] and the fact that we can utilize the networks constructed under the modularity principle with respect to the granularity of information of data. Since choosing a single best method for a given problem like winner-takes-all easily results in losing some useful models.

2 Backgrounds

2.1 Support Vector Machine

The Support vector machine (SVM) produced by Vapnik in 1995 is a method to estimate the function classifying the data into two classes [12]. The basic idea of SVM is to construct a hyperplane as the decision surface in such a way maximizing the margin of separation between positive and negative examples. 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 [13]. Given a labeled set of M training samples (X_i, y_i), where $X_i = \{x_1, x_2, \cdots, x_n\} \in R^N$ and y_i is the associated label,

$y_i \in \{-1, 1\}$, the discriminant hyperplane is defined by:

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

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 α_is (support vectors) [14].

2.2 Self-Organizing Map

The Self-organizing map (SOM) by Kohonen defines a mapping from the input data space onto an output layer by using Kohonen's unsupervised learning algorithm [15, 16]. The SOM has an input layer and an output layer. The output layer consists of N nodes, each of which represents a vector that has the same dimension as the input pattern. For a given input vector X, the winner node m_c is chosen using the Euclidean distance between X and neighbor nodes m_i. The weight vector of winner node is updated by the following formula:

$$\|x - m_c\| = \min_i \|x - m_i\| \tag{2}$$

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\} \tag{3}$$

where $\alpha(t)$ is the learning rate at time t and $n_{ci}(t)$ is a neighborhood function. We have used 10 by 10 map of a rectangular topology and 0.02 of constant learning rate to solve the problem of gender recognition.

2.3 Neural Network Ensembles

According to Osherson *et al.*, 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 [17]. After this notion of modular connectionist systems was first discussed in the mid 1980's by Barto and Hinton, Pollack proposed the cascaded backpropagation architecture [18] and Jacobs developed taxonomy for a class of modular hierarchical connectionist models [19]. Hampshire 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 [20], Lincoln and Skrzypek proposed clustering multiple backpropagation networks [21], and Battiti and Colla suggested the concept of democracy to combine the outputs of different neural network classifiers [22]. These early examples have shown that integrating the multiple modules, often referred as committee machines, could have enhanced the accuracy and generalization capability.

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 [11]. They put Gaussian noise to the original image and performed a 5° of geometric transformation for the input of the hybrid classifier and obtained 93.3% recognition rate (4 errors) of the gender classification over the 60 test sets. In our method, we have examined the recognition ability of neural ensembles with 4 features of same movements, not just from geometrical transformation but from 2D and 3D velocities and positions.

3 Human Movement Data

3.1 Data Acquisition

The movement data were obtained using a 3D position analysis system, called Optotrak, from Northern Digital Inc. 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 styles each. These three types of movements have been chosen as our experimental objects because they are all upper body's movements, simple enough to be played briefly for experiments, have short time duration, take place often around us. Twenty six people participated to build the data set, half of them were males and the others were females. The motion was recorded 10 times repeatedly and 8 of them have been used for the training (totally $1,248 = 26 \times 6 \times 8$) and the rest for the test (totally $312 = 26 \times 6 \times 2$).

3.2 Preprocessing

Each movement has been processed to obtain the start and end points. The start of movement is defined as the instant the tangential velocity of the wrist rises above the 5% of the peak and the end as the instant the velocity passes below the 5% of the peak. The missing data that resulted from a marker going out of camera view have been interpolated to remove this artifact.

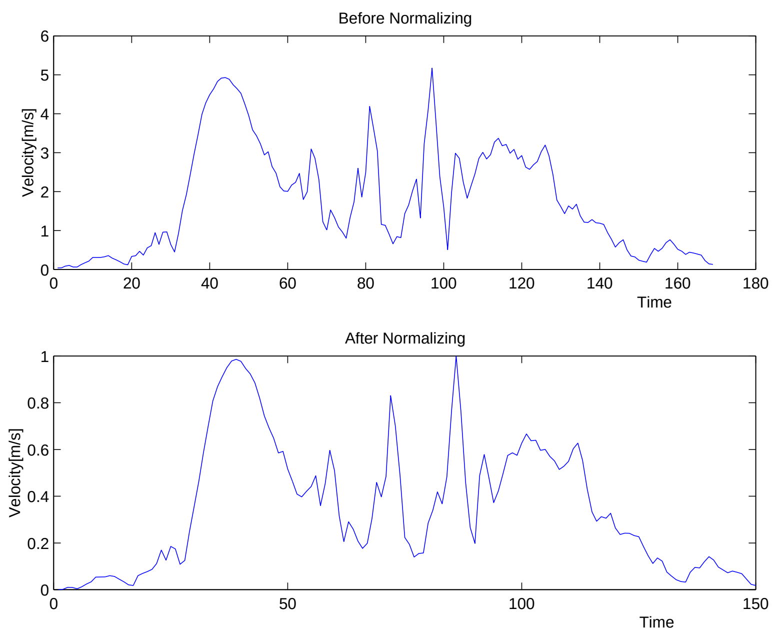

Figure 2: Normalizing a pattern to 150

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.

$$(x_{temple\_1}, y_{temple\_1}, z_{temple\_1}), (x_{shoulder\_1}, y_{shoulder\_1}, z_{shoulder\_1}), (x_{elbow\_1}, y_{elbow\_1}, z_{elbow\_1}), (x_{wrist\_1}, y_{wrist\_1}, z_{wrist\_1}), (x_{finger1\_1}, y_{finger1\_1}, z_{finger1\_1}), (x_{finger2\_1}, y_{finger2\_1}, z_{finger2\_1}), \ldots, (x_{temple\_150}, y_{temple\_150}, z_{temple\_150}), (x_{shoulder\_150}, y_{shoulder\_150}, z_{shoulder\_150}), (x_{elbow\_150}, y_{elbow\_150}, z_{elbow\_150}), (x_{wrist\_150}, y_{wrist\_150}, z_{wrist\_150}), (x_{finger1\_150}, y_{finger1\_150}, z_{finger1\_150}), (x_{finger2\_150}, y_{finger2\_150}, z_{finger2\_150})$$

Figure 3: Input pattern (3Dpos)

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 have 150 in length and be between

0 and 1 in amplitude. As can be seen in Figure 2, even after the normalization, the characteristic of pattern still has been preserved well. Each movement pattern collected from six points has been normalized separately, and the patterns are rearranged in the order shown in Figure 3.

An example of the final set of input patterns has been shown in Figure 4. This is the 3Dpos training set containing 1,248 patterns, each of which has 2,700 dimensions ($6 \times 3 \times 150$). The x axis is the dimension of each vector and y axis is the pattern ID number. We can see there are 3 different types of patterns visually: The first 416 patterns are knocking, the next 416 are waving and the last are lifting movements. Each type of movements has been occurred in neutral and angry styles each.

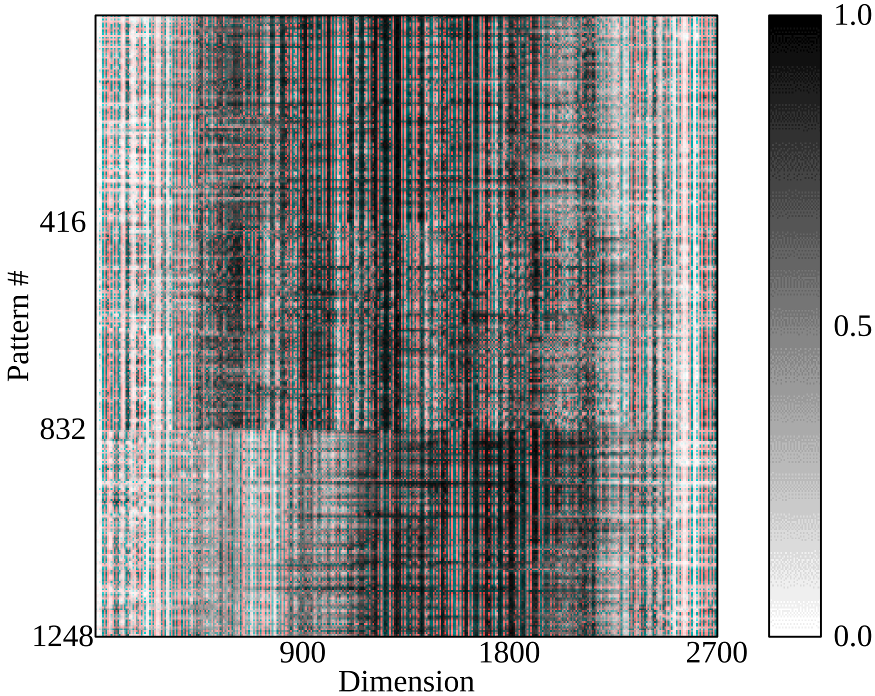

Figure 4: Learning data set space (3Dpos)

4 Neural Ensembles

4.1 Ensemble of Modular MLPs

Multilayer perceptron (MLP) is commonly used in many fields of pattern recognition due to its powerful and stable learning algorithms. The backpropagation based on the delta learning rule is a good example [23, 24, 25]. The power of the backpropagation algorithm lies in two main points: local for updating the synaptic weights and biases of the multilayer perceptron, and efficient for computing all the partial derivatives of the cost function with respect to these free parameters [25]. In a 3-layered MLP, given input vector X, hidden neuron Z and output neuron Y are activated by following equations:

$$Z = f[Net(w_1 X)] \tag{4}$$

$$Y = f[Net(w_2 Z)] \tag{5}$$

where w_1 and w_2 are the weights between first and second, and second and third layers. $Net(.)$ is the weighted sum of nodes connected from the prior layer and $f[.]$ is a sigmoid activation function. The errors on the output and hidden layers (δ_2 and δ_1) are expressed as:

$$\delta_2 = (d - Y) f'[Net(w_2 Z)] \tag{6}$$

$$\delta_1 = \left(\sum \delta_2 w_2\right) f'[Net(w_1 X)] \tag{7}$$

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

$$w_2(t+1) = w_2(t) + \alpha \delta_2 Z \tag{8}$$

$$w_1(t+1) = w_1(t) + \alpha \delta_1 X \tag{9}$$

where α is a learning constant. These procedures are iterated until the network reaches on a certain level of recognition rate or the given maximum number of iterations.

The ensemble of modular MLPs called EMMLP proposed in this paper is shown in Figure 5. Since there are six types of movements, we have divided the whole data into six sub classes to reflect the locality of our data. The EMMLP consists of a motion classifier that has been trained to recognize data in the corresponding sub class, and six modular MLPs that are in charge of each of sub classes. Once the given input pattern is assigned to a certain class, the modular MLP makes the decision of classification. All modular networks have been trained separately only with the data that they are in charge of. EMMLP can perform the best if the motion classifier gives perfect answer, but it possibly takes advantages that even misclassified patterns by the motion classifier could be correctly classified by chance.

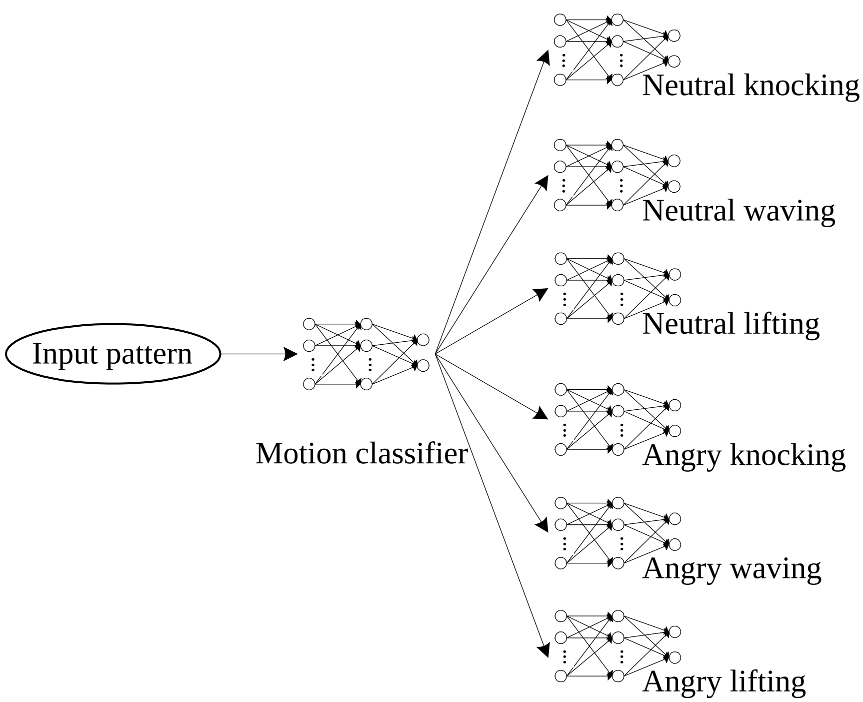

Figure 5: Ensemble of modular MLPs

Each MLPs have 900 (For 2Dvel and 3Dvel, every six joints produces one velocity value: $6 \times 1 \times 150$), 1800 (For 2Dpos, two values are coming out, y and z, from every joints: $6 \times 2 \times 150$) or 2700 (For 3Dpos, three numbers coming out, x, y and z: $6 \times 3 \times 150$) input nodes, and 50 hidden and 2 output nodes. The learning rate and momentum are chosen empirically from 0.001 to 0.1, and they are different for the different networks in the ensemble so that each network performs as an expert for his task. All the networks are trained until the recognition rate on the training data reaches 98.0%. "Male" is encoded as [1 0] and "female" as [0 1] for the output layer.

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 [26]. 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 entropy-based criterion that has been used for the selection of the attribute is called the gain ratio criterion [27].

Let X be a possible test (attribute selection) that partitions the training set T into n sub sets $(T_1, T_2, \cdots, T_n)$. When $|T|$ is the size of T and $info(T)$ is the expected information (entropy) that T conveys, $split\_info(X)$, the potential information obtained by partitionning a set of cases, and $gain\_ratio(X)$ are expressed as [26, 27]:

$$info(X,T) = -\sum \left(\frac{|T_i|}{|T|}\right) \log_2 (T_i) \tag{10}$$

$$gain(X) = info(T) - info(X,T) \tag{11}$$

$$split\_info(X,T) = -\sum \left(\frac{|T_i|}{|T|}\right) \log_2 \left(\frac{|T_i|}{|T|}\right) \tag{12}$$

$$gain\_ratio(X) = \frac{gain(X)}{split\_info(X)} \tag{13}$$

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

The second classifier, ensemble of modular MLPs/decision tree (EMMLP/DT) is as shown in Figure 6. We have used 4 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, each of the 4 MMLPs produces 6 outcomes, and finally the output vector $(x_1,\, x_2,\, \cdots\, ,\, x_{24})$ has been chosen for training, which is the input to the decision tree.

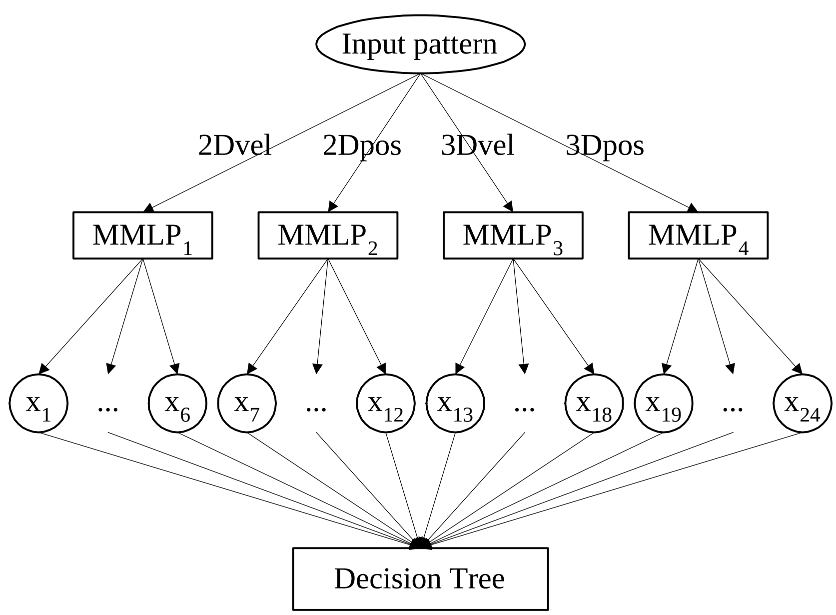

Figure 6: Ensemble of modular MLPs and decision tree

The decision tree consists of decision nodes and leaves. Leaves tell us the class label, indicating "male" or "female," and the decision nodes specify a test to be carried out on a single attribute value,

Table 1: Recognition rate of classifiers [%]

Classifier	Data Representation				Average
	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

with one branch for each possible outcome of the test. A person's movement is an object, and the attributes are the dimensions of input vectors that are continuous numbers normalized between 0 and 1.

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 Experimental Results

5.1 Recognition Rates

The final result of recognition rates with respect to the classifiers and the data representations has shown in Table 1. Experiments with ensembles have been repeated 3 times. On the average, the EMMLP produced the best performance and the EMMLP/DT (standard deviation: 1.7) was the second, followed by MLP, SVM, SOM, DT and human participants in order. The performances of classifiers vary depending on the representation used. The recognition rate of EMMLP varies from 80.0% to 86.5%. Since the EMMLP/DT uses all representations as its input 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.

For the comparison with human performance, Pollick *et al.* have conducted experiments with human [28]. They also use the dynamic point light displays to train to discriminate the people's gender. Human has been trained with a process of "pre-test, training and post-test," with the same training and test dataset that we have used. The human performance is just above 50% of recognition. This means human may not have a principle or criterion enough to discriminate the gender. The reason might be that we have excluded the use of clues such as size or shape of objects as much as possible, which are subject to be used by people to perceive objects. However, since the point light display represents the biological motions as only moving spots, it is hard for the human to recognize the actors' gender.

Figure 7 shows the decision tree constructed by EMMLP/DT method. The left branches of each decision node 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 where x_{18} is larger than 0.48 and x_{17} is smaller than 0.003 would be classified as a female's movement. Generally, machine learning classifiers performed much better than human. While the EMMLP yields up to 86.5% of recognition rate, humans produce just 51.3%.

There is a trade-off between the accuracy and the computational time required. It is known backpropagation neural network learns well but it takes time because MLP updates the weights and

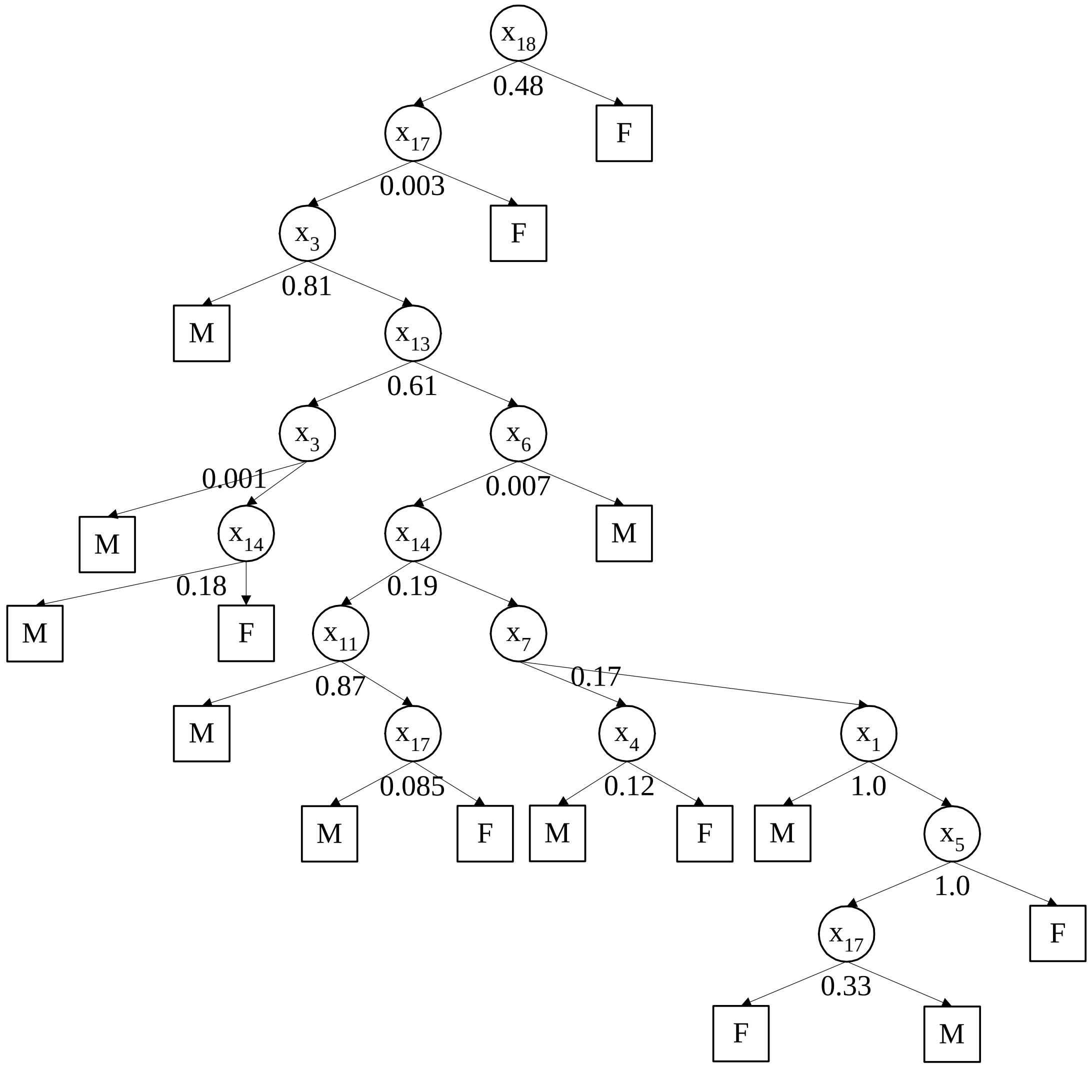

Figure 7: Result of EMMLP/DT

biases iteratively until many epochs are passed. The running time of classifiers were about DT < {SVM, SOM} << MLP << EMMLP << EMMLP/DT. For the EMMLP, when the running time of MLP is t, motion classifier may also have t, and six small classifiers $1/6 \times t$ per each because small classifiers have been trained and tested within their 1/6-subsets of train and test data, thereby, $2t$ $(t + 1/6 \times t \times 6)$ is required. However, EMMLP/DT has 24 small classifiers and DT, so that at least $24 \times 1/6 \times t + a$ (a is the running time of DT).

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 as in Table 2.

Based on the numbers from the confusion matrix, FalseAlarm_rate and Hit_rate are calculated as below:

$$\text{Hit\_rate} = \frac{\text{Hit}}{\text{Hit} + \text{Miss}} \tag{14}$$

$$\text{FalseAlarm\_rate} = \frac{\text{False alarm}}{\text{False alarm} + \text{Correct rejection}} \tag{15}$$

Table 2: Four possible cases

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

We can draw the internal response probability distribution curves (usually normal distribution, but this can be Poisson's distribution etc. if one has the prior knowledge on this) of his observation as in Figure 8. The discriminability can be thought of the distance between two peaks. If the seperation between two distribution gets bigger (shifting the signal+noise curve to the right) and the decision criteria stay on the same level, then Hit_rate will increase, whereas the FalseAlarm_rate decrease. Consequently, the better detector (classifier) will have the higher d'.

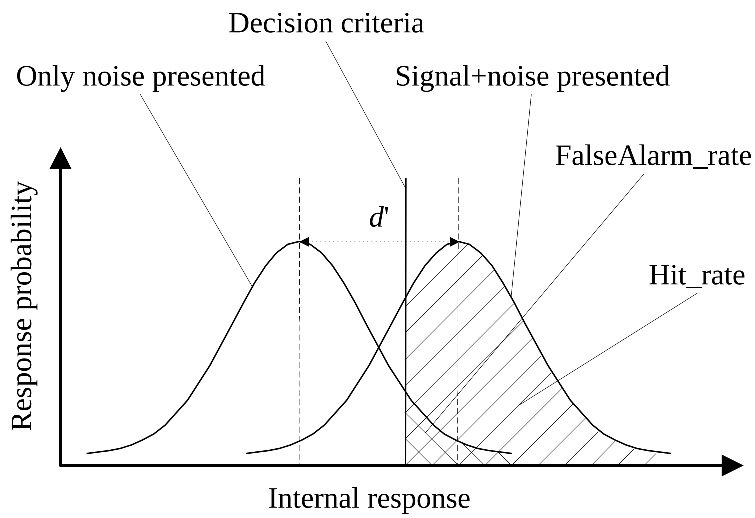

Figure 8: Definition of discriminability

If there exist observer 1 and 2 and an ideal observer whose discriminabilities are $d_1{}'$, $d_2{}'$ and $d_{ideal}{}'$, respectively, Barlow defined the efficiency of observer 1 (E_1) and 2 (E_2) as [29]:

$$E_1 = \left(\frac{d_1{}'}{d_{ideal}{}'}\right)^2, E_2 = \left(\frac{d_2{}'}{d_{ideal}{}'}\right)^2 \tag{16}$$

However, since we do not know the ideal observer, extending this formula to obtain the efficiency ratio of observer 1 over observer 2 leads to:

$$ER_1 = \frac{E_1}{E_2} = \frac{\left(\frac{d_1{}'}{d_{ideal}{}'}\right)^2}{\left(\frac{d_2{}'}{d_{ideal}{}'}\right)^2} = \left(\frac{d_1{}'}{d_2{}'}\right)^2 \tag{17}$$

This tells how more efficient observer 1 is utilizing detection information behind the input signal then observer 2. The efficiency ratios of ensemble classifiers over human subjects have been calculated. We have assumed that 'signal' is male's and 'noise' is female's pattern.

Table 3 is the result of discriminability. According to the average, the EMMLP produces the best result and the EMMLP/DT is the second. Similarly to the recognition rate results, the 3Dpos is the most effective representation. The EMMLP with 2Dpos produces the highest value of d', 1.30. In contrast to the neural networks, human subjects were unable to judge gender from arm movements

Table 3: Discriminability of classifiers

Classifier	Data Representation				Average
	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 4: Efficiency ratio of ensemble classifiers

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

although previous literature has shown that structural differences exist in such movements. It might be because we have excluded such familiar clues as size or shape of objects when one discriminating gender.

The efficiency ratio of ensemble classifiers against the human participants is as shown in Table 4. The EMMLP with 2Dpos yields the highest efficiency ratio of 100.0, which is simply because the EMMLP with 2Dpos has the biggest number of hits and correct rejection. It seems that the human recognition is very poor. In fact, it is lower than what Johanssen's experiment. It might be because we have excluded such clues as size or shape of objects, which people might feel much familiar with when discriminating gender. Also, Johanssen have used ten point lights attached throughout the whole body to observe human gait, but we used only six point lights on the right side of the upper body. Additionally, the object movements that we have used is more restricted than Johanssen's (the degree of freedom of arm movements is much smaller than that of walking), so that people may get more difficulty to obtain the categorical features from the motions.

6 Concluding Remarks

In this paper, to classify the human's gender, we have proposed two ensemble classifiers called EMMLP and EMMLP/DT and compared the performance with other conventional classifiers. As the result, the EMMLP has performed the best and the EMMLP/DT is the second both in recognition rate and discriminability. Even though the simple MLP and DT yield approximately 75% and 70% of recognition rate, we have enhanced the performance up to 86.5% by combining modular MLPs with ensemble of the modular MLPs and decision tree. Among representations, 3Dpos is the best feature in most of the classifiers. The results of discriminability and efficiency ratio show that the neural ensembles are much better to recognize the actor's gender than non-ensemble neural networks as well as human subjects.

References

[1] W Wolff, *The Expression of Personality: Experimental Depth Psychology*, Harper & Brothers, NY, 1943.

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

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

[4] S Ullman, *The Interpretation of Structure from Motion*, Cambridge, MIT Press, MA and London, 1979.

[5] R Rashid, "Towards a system for the interpretation of moving light displays," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 574–581, 1980.

[6] ST Barnard, WB Thompson, "Disparity analysis of images," *Technical Report 79-1*, University of Minnesota, Minneapolis, 1979.

[7] JE Cutting, LT Kozlowski, "Recognising friends by their walk: Gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.

[8] J Ryu, S-B Cho, "Searching for optimal features for gender recognition with neural network classifier," *Proceedings of International Conference on Soft Computing*, pp. 590–596, Iizuka, 2000.

[9] RP Lippman, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4–22, 1987.

[10] S Haykin, *Neural Networks*, 2nd Ed., Prentice Hall, pp. 351–391, 1999.

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

[12] VN Vapnik, *The Nature of Statistical Learning Theory*, Springer, NY, 1995.

[13] V Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, NY, 1982.

[14] B Moghaddam, MH Yang, "Gender classification with support vector machines," *Proceedings of 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 306–311, 2000.

[15] T Kohonen, "The self-organizing map," *Proceedings of IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[16] T Kohonen, *Self-Organizing Maps*, Springer, Berlin Heidelberg, 1995.

[17] DN Osherson, S Weinstein, M Stoli, "Modular learning," *Computational Neuroscience*, EL Schwartz (Ed.), pp. 369–377, Cambridge, MIT Press, MA, 1990.

[18] J Pollack, "Cascaded back-propagation on dynamic connectionist networks," *Proceedings of Ninth Annual Conference Cognitive Science Society*, pp. 391–404, 1987.

[19] R Jacobs, "Initial experiments on constructing domains of expertise and hierarchies in connectionist systems," *Proceedings of 1988 Connectionist Models Summer School*, San Mateo, CA, pp. 144–153, 1988.

[20] JB Hampshire, A Waibel, "The Meta-Pi network: Building distributed knowledge representations for robust multisource pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 751–769, 1992.

[21] WP Lincoln, J Skrzypek, "Synergy of clustering multiple backpropagation networks," *Advances in Neural Information Processing Systems*, D. S. Touretzky (Ed.), San Mateo, Morgan Kauffmann, CA, vol. 1, pp. 650–657, 1990.

[22] R Battiti, AM Colla, "Democracy in neural nets: Voting schemes for classification," *Neural Networks*, vol. 7, pp. 691–707, 1994.

[23] BVS Murthy, "Delta learning law for a single neuron," *IEEE International Joint Conference on Neural Networks,* vol. 3, pp. 1779–1782, 1999.

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

[25] HD Beale, *Neural Network Design*, vol. 11, pp. 1–47, PWS Publish Company, 1996.

[26] JR Quinlan, "The effect of noise on concept learning," *Machine Learning: An Artificial Intelligence Approach*, RS Michalski, JG Carbonell, TM Mitchell (Eds.), San Mateo, Morgan Kauffmann, CA, vol. 2, pp. 149–166, 1986.

[27] JR Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 2, pp. 77–90, 1996.

[28] FE Pollick, V Lestou, J Ryu, S-B Cho, "Estimating efficiency in the categorization of biological motion," *Abstracts of Vision ScienceS Society First Annual Meeting*, pp. 31, 2001.

[29] HB Barlow, "The efficiency of detecting changes in random dot patterns," *Vision Research*, vol. 18, pp. 637–650, 1978.