Fusion of Speech and Face by Enhanced Modular Neural Networks

Rahul Kala[1], Harsh Vazirani[2], Anupam Shukla[3], Ritu Tiwari[4]

[1, 2, 3,4] Department of Information Technology, Indian Institute of Information Technology and Management Gwalior, Gwalior, MP, INDIA
[1]rahulkalaiiitm@yahoo.co.in, [2]harshiiitmg@gmail.com, [3]dranupamshukla@gmail.com, [4]rt_twr@yahoo.co.in

Citation: R. Kala, H. Vazirani, A. Shukla, R. Tiwari (2010) Fusion of Speech and Face by Enhanced Modular Neural Network, *Proceedings of the Springer International Conference on Information Systems, Technology and Management*, Bankok, Thailand, pp 363-372.

Final Version Available At: http://link.springer.com/chapter/10.1007%2F978-3-642-12035-0_37?LI=true

Abstract. Biometric Identification is a very old field where we try to identify people by their biometric identities. The field shifted to bi-modal systems where more than one modality was used for the identification purposes. The bi-modal systems face problem related to high dimensionality that may many times result in problems. The individual modules already have large dimensionality. Their fusion adds up the dimensionality resulting in still larger dimensionality. In this paper we solve these problems by the introduction of modularity at these attributes. Here we divide various attributes among various modules of the modular neural network. This limits their dimensionality without much loss in information. The integrator collects the probabilities of the occurrences of the various classes as outputs from these neural networks. The integrator averages these probabilities from the various modules to get the final probability of the occurrence of each class. This averaging is performed on the basis of the efficiencies of the modules at the time of training. A module that is well trained is hence expected to give a better performance than the one which is not well trained. In this manner the final probability vector may be calculated. Then the integrator selects the class that has the highest probability of occurrence. This class is returned as the output class. We tested this algorithm over the fusion of face and speech. The algorithm gave good recognition of 97.5%. This shows the efficiency of the algorithm.

Keywords: Modular Neural Networks, Artificial Neural Networks, Fusion Methods, Classification, Speaker Recognition, Face Recognition, Ensemble

1 Introduction

Biometric Identification deals with the identification of people by their biometric identities [1, 2]. There are numerous biometric identities that have been very commonly used. These techniques are usually divided into two categories. These are physiological biometrics that deals with static characteristics. These include face [3], ear, iris, hand geometry, etc. The other category is the behavioral biometrics that includes the active features such as speech, signature, handwriting etc. The limited recognition efficiencies of these systems resulted in a shift towards the multi-modal or bi-modal recognition systems [4, 5, 6, 7]. Here the motivation was to mix two or more modalities in order to attain even greater efficiencies. The problems or limitation of one modality could be solved by the other modality. This resulted in fusion of known biometric modalities that include fusion of face and speech [8, 9, 10, 11, 12, 13], fusion of face and ear, fusion of iris and face, face and fingerprint etc. Multi-modal systems with which combine more modalities was also developed. The fused methods employed the mixing up of the attributes from both the biometric identities and giving the same to the recognition systems.

One of the major problems with the biometric identification systems is that of high dimensionality. The extremely large dimensionality of these problems results in the use of good feature extraction systems. In many problems like face recognition, the dimensionality is still large. This may especially have a problem if the size of data in the training data set is too large. This would normally make the system very slow in learning. Also the error at the time of training would be comparatively large. The number of epochs has to be limited. This results in poor performance of these systems.

The bi-modal systems are formed by the fusion of attributes [4, 5, 6, 7]. As a result the dimensionality associated with the fused systems would be still very large. This large dimensionality would result in the same set of problems to an even larger extent. This may make these systems unworkable whenever the size of data set is very large. This induces a big limitation in these systems and restricts their performance.

The Modular Neural Networks (MNN) is the solutions to these types of problems [15, 16]. The basic principle of these networks is to make use of modularity in the system. These divide the entire problem into various modules. Then all the modules calculate the solution to their part of the problem. This happens in parallel between the various modules. The solution so generated is computed to an integrator. The integrator does the task of integration of the results that come from the various modules. It fetches the individual results from each of the modules and then uses these results to calculate the final output that is returned as the output of the system.

There have been numerous ways and models of the ANN. One of the most commonly used models in classification systems is the ensemble [17]. The ensemble architecture is the same as that of the MNN. It divides the problem into various modules and then integrates them by an integrator. In ensemble there is a poling mechanism that is employed to carry out the classification. Each of the module votes for some class. These votes are added and the final class that wins is declared as the winner.

The MNNs developed so far try to employ a hierarchical approach where the input is mapped to some network that performs the task or a group of networks that perform the task which is later of integrated. This would not solve the problems arising out of dimensionality because of the fact that the dimensionality in these systems remains the same. Also it is further not possible to use any dimensionality reduction technique, as this would result in reduced system performance and loss of information. Hence we need better mechanisms for the application of modularity.

In this paper we propose the modularity to be applied at the various attributes. This means that we intend to distribute the attributes among the various modules for the task of identification. This division needs to be done judiciously so that the recognition is good.

Another problem with the ensemble is that the various modules can only return one particular class as their output. This would mean the taking of wrong decisions at various times when a number of modules cannot decide output between some classes. In our approach every module returns a set of probabilities of the occurrence of every class. This gives a lot of information to the ensemble for the combination of the various classes. This system, as against the polling mechanism, may be viewed as a system where a small set of experts sit together and discuss the final output class, in place of just a voting which is not a good solution in case the number of experts are limited.

The novelty of the paper lays four fold. (i) Firstly we modify the present ensemble approach to make it better for the classification problems. This is done by the introduction of probabilities (ii) Secondly we introduce the concept of modularity at the attribute level that would help the system in faster training and dimensionality control along with the performance boost. (iii) Thirdly we suggest the mechanism of weighing of the various modules that may further help in improvisation of the results by eliminating the bad modules from over affecting the decisions (iv) Fourthly this is applied to the problem of fusion of face and speech for the task of person identification for achieving even greater performance. Here we reduce dimensionality of the problem without loss of information.

The innovation lies in the selection of the number of modules and their attributes. This needs to be done in such a way that all attributes get covered and each module is easily able to solve the problem with good efficiencies even if operating alone.

This paper is organized as follows. Section 2 talks about the general recognition systems and various attributes that are extracted for the problem for both the speech as well as the face. Section 3 presents a native fusion algorithm between face and speech. Section 4 would present the entire algorithmic framework. Finally in section 5 we present the results and in section 6 we present the conclusions.

2 Recognition System

Any recognition system involves various stages. The final output is the recognized person or identity. Here the first task is the data collection that acquires the data in the system. In the problem of fusion of face and speech, the camera is used to take the photograph of the person. At the same time the microphone may be used to capture

his voice. Ease of the user is a major criterion that needs to be taken care of. Here the system would be very simple to use for the user where the image and speech can be acquired simultaneously.

The next step comes is the image preprocessing. This is needed for the noise removal as well as to highlight the features. In case of the face the input is in the form of image that requires the application of noise removal operators and binarization. In case of speech the input is a signal that may be freed from noise by the application of noise removal filters.

The next task is segmentation. Here we segment the image and the features. In image the task is concerned with application of gradient mask, dialization, filling up of holes, etc. In speech we segment each and every word of the spoken sentence.

Then feature extraction is done. Here we extract the features for dimensionality reduction. The extracted features must be such that they lead to large inter-class distances and small intra-class distances. They must be relatively constant when the same face is clicked numerous time, or the person speaks various times. The features used for the speech and face are discussed below.

For the speech we extract a total of 11 features. These are time duration, number of zero crossing, max cepstral, average PSD, pitch amplitude, pitch frequency, peak PSD and 4 formants (F1-F4). All these features are extracted using the signal processing toolbox of MATLAB and the inbuilt MATLAB functions. These features are all widely used in research for the speech and speaker recognition and verification systems. They are found to remain stable each time the same person is recorded and analyzed.

For the face there were a total of 13 features extracted [18]. These are length of the eye 1, width of the eye 1, center dimension x 1, center dimension y 1, length of the eye 2, width of the eye 2, center dimension x 2, center dimension y 2, length of the mouth, width of the mouth, center dimension x, center dimension y, distance between eye and eye, distance between eye and mouth.

These features as well are extensively used in research related to face recognition and verification. In case of face the problems of the stability of the features is even more important as it is largely dependent on light, expressions and other variations that are possible. These features are relatively more stable.

3 Native Fusion Algorithm

After these features, the first task done was the fusion of the features so collected into a fused system. This is motivated directly from literature where such systems are found to be giving better performances and results as compared to the native methods. We solve the problem in 2 cases. In the first the entire feature vector was used directly as an input to the problem. In the second approach we manually restricted the inputs before giving them as the input.

As the first case we used the entire feature vector of both the systems combined as an input to the system. This included the extracted 11 features from speech and 13 features from face. This made a total of 24 inputs to the classifier for the recognition

systems. Naturally the classifier took a lot of time to train. This was the classical approach of fusion of the face and speech that we applied.

In this case we manually selected some features from the speech and some from the face and only these were allowed to be used for the recognition system. This limited the number of features that were used and hence resulted in a better training time and training efficiency. However there was a permanent loss of information as many of the extracted features were not at all used by the system. This may many times lead to reduced efficiencies.

The features that we selected for this problem are Formant Frequency F1 to F4, Peak & Average Power Spectral Density, Length and width of the Mouth, Distance between Center Points of Eye 1 & Eye 2, Distance between Center Points of Eyes & Mouth. These made a total of 7 attributes for the system.

4 Algorithm

The overall algorithm is built over the modular neural network approach. This approach believes in the principles of modularity of the problem where the entire problem is first divided into modules and then solved independently by each module. Once each module returns its results, an integrator is used for the task of integrating the solutions of the various modules and calculating and returning the final output vector.

In this problem of fusion of face and speech, we first divide the problem into modules. Here the modules are divided into modules. Every module gets a set of attributes. It is possible that an attribute is given to more than one module. Similarly it is possible that an attribute is not given to any of the attributes at all. This however must be avoided as it leads to a loss of valuable information.

Then each module independently calculates the output. These use ANNs for the task. At the end each ANN returns the probabilities of the occurrence of each of the class. These probabilities lie in the range of -1 to 1. An occurrence of -1 means that the given class is completely absent and vice versa. The numbers denote the certainty of the ANN in the occurrences of the class as the final output class.

At the end the integrator does the task of combining the individual solutions to give the final output. This is in the form of 1^{st} averaging the various probabilities returned by the individual systems. Here the performance of the systems at the time of training is used as weights. Then the summation takes place to return the final class that is declared as the final output. The basic working of the system is shown in figure 1. We discuss each of the steps in the next section.

4.1 Modules

Here we are supposed to exploit the modularity in the features. The basic motive is to ensure that each module after getting its feature set must be in a condition to appreciably solve the problem. It must have the related attributes to enable it to do so. Hence the attributes given to any module must be diverse and must collectively supply the entire information. Also loading too many features to an ANN would be

not desirable. Here we also try to ensure that all attributes collectively get the complete feature set. This would avoid the loss of information from the system.

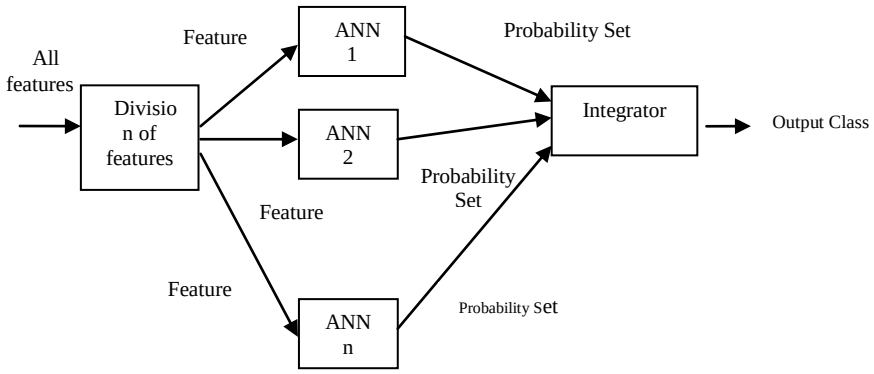

Figure 1: The general structure of the algorithm

In this approach we keep the speech and face features completely different. We then divide the speech attributes into two parts and similarly the face attribute into two parts. In this system we select some speech features to be used as inputs for the first module. These are time duration, max cepstral, pitch amplitude, peak PSD, F2 and F4. The second module contains the rest of the speech features. These are number of zero crossing, average PSD, pitch frequency, F1 and F3.

The last two modules are to cover the facial features. Here also we follow a similar technique. The 3rd module covers the features length of the eye 1, width of the eye 1, center dimension x 1, center dimension y 1, length of the mouth, distance between eye and eye. The rest of the facial features belong to the fourth module. This includes length of the eye 2, width of the eye 2, center dimension x 2, center dimension y 2, width of the mouth, center dimension x, center dimension y, distance between eye and mouth. The structure in general related to the division of the features among modules is shown in figure 2.

4.2 Artificial Neural Networks

The job of classification of the inputs is carried out with the help of ANNs with BPA as the training algorithm. The ANNs are a natural choice because of their ability to learn from the historical data and to generalize the results. The ANNs map any input to some class or person here. We use a classificatory model of ANN here. This has as many output neurons as the number of classes. Each output neuron stands for some person or class. The output at this neuron for any input i is the probability of occurrence of this person or class according to the ANN. Hence the ANN gives as its output c number of probabilities in the output vector. Let this output vector for any input i be represented by $<v_{i1}, v_{i2}, v_{i3}, \ldots v_{ic}>$.

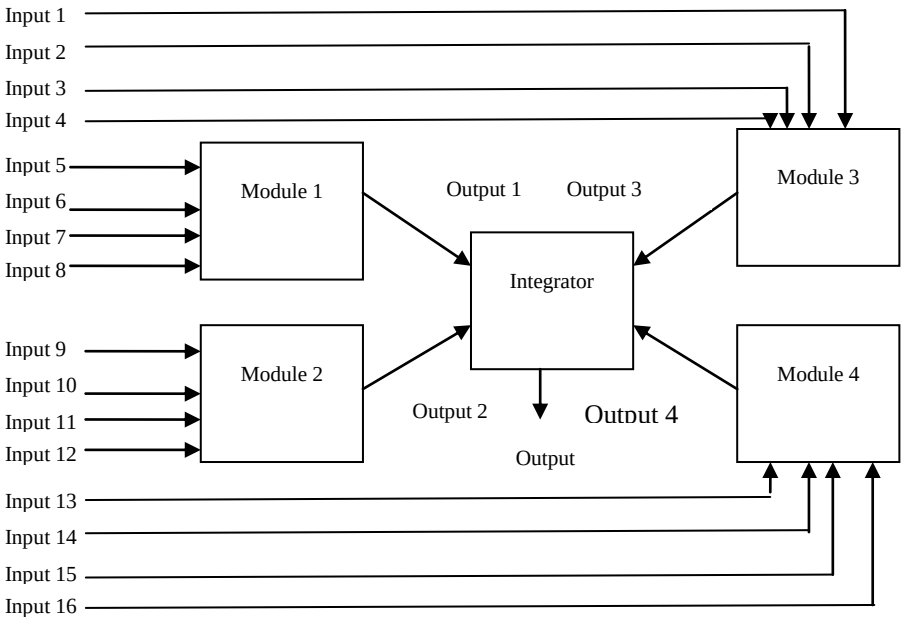

Figure 2: The division of inputs between modules

The probabilities here are measured in the range of -1 to 1. A probability of 1 means a certainty of the output class with full confidence. On the other hand a probability of -1 means that according to the ANN this class is not the output with full confidence. Hence the ideal output for any input for the ANN should be a 1 for any one element of the output vector and a -1 for all the other elements. However due to practical reasons, the output lies anywhere in the complete region.

4.3 Integrator

The last part to implement according to the entire algorithm is the integrator. This is the major part of the whole system that does the task of finding the final output class after getting inputs from the individual modules. The integrator analyzes all the outputs by the various modules and then decides the final output class that according to the system is the output.

The input given to the integrator is the solution vector of every module of ANN. Let the vector of the module i be $<v_{i1}, v_{i2}, v_{i3}, \ldots v_{ic}>$ where each v_{ik} is the probability of the occurrence of the class k measured on a scale of -1 to 1. The integrator decides the output by first taking the weighted averages of the probabilities given to it and then selecting the class with the maximum probability. This class is declared as the winner. This is shown in figure 3.

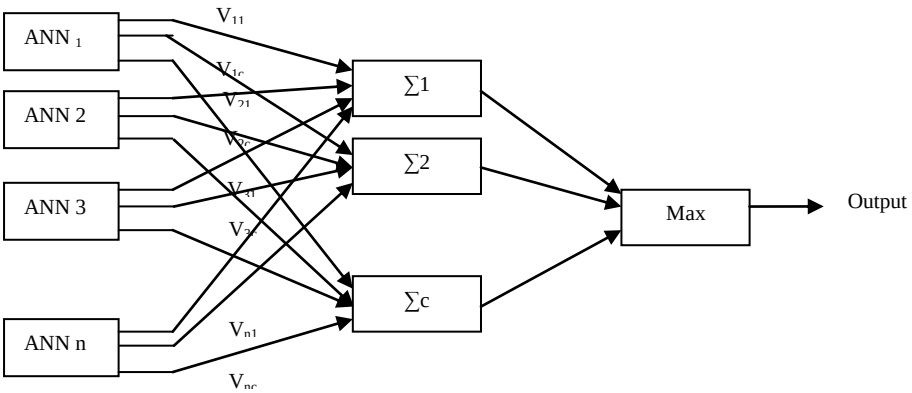

Figure 3: The Integrator

The weighted average is calculated for each and every class in the system on which the output can map to. The weights of the various modules or ANNs here are their performances on the training data. Each ANN was given the same data set for the training purposes with a different feature set. The performance here is calculated as a ratio between the numbers of elements the ANN correctly classifies in the validation data set by the total number of elements in the validation data set. The higher the performance of the ANN in the validation data set, the more would be its weight and more dominant it would be to decide the final output at the time of integration.

Using these calculated weights, the integrator calculates the weighted average for all the classes. This gives the final probability vector comprising of c probabilities with each probability associated with some class. These probabilities again lie in the range of -1 to 1 with the same meaning of the probabilities.

The next task of the integrator is to find out the final output class. For this the integrator selects the output class with the largest value of the probability out of all the available classes. The class or the person corresponding to this largest probability is declared as the winner and this is the output that the system gives.

5 Results

The testing of the algorithm was done by experimenting and validating the results using a self made database. The database consisted of data of 20 people whole pictures had been taken multiple times. At the same time their voices had been recorded with same words multiple times. This made a system which had to identify the person out of the data of 20 people available in the database.

The features of both face and speech were recorded using MATLAB tools and functions. This transformed the whole data into numerical forms. Normalization was carried out for each of the attributes to make it lie in the range of 0 to 1.

The first experiment conducted was on face and speech separately. We tried to test the performance of each of these modalities without the presence of the other. ANN with BPA was used for the classification purposes. The efficiency received in this experiment over the database was 90.0% for the speech and 92.5% for the face recognition system. This was high enough but necessitated the need of better system.

Then all the features were divided into 4 sets or modules using the strategy discussed earlier. This division into 4 modules resulted in the system being more modular in nature. Each of the ANN was trained with BPA. The training because of the limited dimensionality happened much more early as compared to the other fused system. Further the training reached good performances and low error rates.

The testing was done using the system so developed. Using this system over the built database, we received an accuracy of 97.5% which is much larger than the two systems developed and tested separately.

6 Conclusions

In this paper we took the problem of fusion of face and speech. Using this problem as a means we studied a good method of reducing dimensionality in the problem that was causing effects to the performance of the system. For this we developed an algorithm based on the MNN. The algorithm introduced modularity in the features and divided them into various modules. Each module could be separately used as a system of its own in the classificatory problem with just a little reduced performance.

Then each of these modules was given its share of every input and this in turn returned the probability vector where each output denoted the probability of occurrence of some class. These were combined by using weighted average by the integrator. The integrator then selected the class with the maximum probability of occurrence and this was declared as the final output of the system.

The system so developed has various functionalities that are better than the original fused methods. This solves the problem of high dimensionality that is prevailing in the original methods. This results in better training and training in reduced time and reaching of larger number of epochs. This has a good effect on the system performance and we are able to reach a higher level of accuracies.

The algorithm also proposes a change in the ensemble manner of poling. Here we proposed a probability based poling that can be better than the poling used by the ensemble. Also the weighing of the various modules is carried out which gives a better performance and reduces the influence of bad modules. This further increases the system performance.

Another innovation in this algorithm is the division of attributes between the modules. This division is carried out in a manner to allow maximum efficiencies of each module. This is done by reducing dependencies and giving diverse inputs. This may be generalized to any classificatory problem in general.

The system so obtained gave a good recognition of 97.5% over the self made database and the self extracted features. This is highly encouraging and proves the efficiency of the algorithm as well as the views presented. This was much higher than the performance of the same database with single modalities.

Even though we have proved the working of this algorithm and received good results, a lot may be done in the future. The algorithm needs to be worked upon large databases that would fully exploit and justify the scalability factors of the algorithm. Further we used a set of features for the face and speech. The work over the other possible features may be done. Also different feature combinations may be tried in future. The use of the same algorithm may be done for other multi-modal systems as well. This may even be generalized to any classificatory problem. The weighing factor we discussed was over the validation data set performances. Possibilities of the other weighing strategies may be tried in future and their results may be compared.

Acknowledgments. This work is sponsored and supported by Indian Institute of Information Technology and Management Gwalior.

References

1. Souheil Ben-Yacoub, Yousri Abdeljaoued & Eddy Mayoraz, Fusion of Face and Speech Data for Person Identity Verification, IEEE Transactions On Neural Networks, Vol 10, No 5, 1065, September 1999
2. Anil Jain, Lin Hong, Sharath Pankanti, Biometric Identification, Communications of the ACM, Volume 43, Issue 2 (February 2000), pp 90 – 98
3. Ching-Han Chen, Chia-Te Chu, Combining Multiple Features for High Performance Face Recognition System, International Computer Symposium(ICS2004) Taipei, (Dec 2004), pp.387-392.
4. Robert Snelick, Mike Indovina, James Yen, Alan Mink, Multimodal Biometrics: Issues in Design and Testing, ICMI'03, Canada, (Nov 5-7. 2003), pp.68-72.
5. A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognition Letters, (2003), (24), 2115- 2125.
6. Andrew L. Rukhin, Igor Malioutov, Fusion of Biometric Algorithm in the Recognition Problem, Pattern Recogition Letters, (2001), 299-314.
7. R.W. Frischholz and U.Dieckmann, Bioid: A Multimodal Biometric Identification System, IEEE Computer, (Feb 2000), (33), 64–68.
8. Big¨un, J Big¨un, B Duc & S Fischer, Expert conciliation for multi modal person authentication systems by Bayesian statistics, in Proc 1st Int Conf Audio- Video-Based Biometric Person Authentication AVBPA'97. Berlin, Germany: Springer-Verlag, Lecture Notes in Computer Science, pp 291–300, 1997.
9. T Choudhury, B Clarkson, T Jebara & A Pentland, Multimodal person recognition using unconstrained audio and video, in Proc 2ndInt Conf Audio-Video Based Person Authentication, Washington, DC, pp 176–180, Mar 22–23, 1999.
10. S Ben-Yacoub, Multimodal data fusion for person authentication using SVM, in Proc 2nd Int Conf Audio-Video Based Biometric Person Authentication, Washington, DC, pp 25–30, Mar 22–23, 1999.
11. E K Patterson, S Gurbuz, Z Tufekci & J N Gowdy, Noise-based audio-visual fusion for robust speech recognition, in International Conference on Auditory-Visual Speech Processing, Denmark, 2001.
12. C.Sanderson and K.K. Paliwal, Information Fusion and Person Verification Using Speech & Face Information, IDIAP, Martigny, Research Report, (2002), 02-33.
13. Anupam Shukla, Ritu Tiwari, A Novel Approach of Speaker Authentication by Fusion of Speech and Image Features using ANN, International Journal of Information and Communication Technology (IJICT). Inderscience Publishers, (2008), (1)(2), 159-170.

14. A K Jain, L Hong & Y Kulkarni, A multimodal biometric system using fingerprints, face and speech, in Proc 2nd Int Conf Audio-Video Based Biometric Person Authentication, Washington, D.C., pp 182–187, Mar 22–23, 1999.
15. J Kittler, M Hatef, R P W Duin & J Matas, On combining classifiers, IEEE Trans Pattern Anal Machine Intell., vol 20, pp 226–239, 1998.
16. Fogelman Soulie F., Viennet E., Lamy B., "Multi-modular neural network architectures: applications in optical character and human face recognition", International journal of pattern recognition and artificial intelligence, 1993, vol. 7, no 4, pp. 721-755
17. M.P. Perrone, and L.N. Cooper, "When Networks Disagree: Ensemble Methods for Hybird Neural Networks," Neural Networks for Speech and Image Processing, 1993.
18. Rafael C Gonzalez & Richard E Wood: Digital Image Processing, Pearson Education Asia, 2002.