Farsi Handwritten Recognition Using Combining Neural Networks Based on Stacked Generalization

Reza Ebrahinpour[1], Mona Amini[2], and Fatemeh Sharifizadehi[3]

[1]Brain and Intelligent Systems Lab, Department of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University Lavizan, Tehran, Iran.
[2]Mechatronic Engineering in Department of Mechatronics Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran.
[3]Department of Mathematics and Computer Science, Shahid Bahonar University of Kerman, Iran
ebrahimpour@ipm.ir, Mona.amini.mecha86@gmailcom, sharifizade@ut.ac.ir

Abstract: Stack Generalization is a general method for combining low-level classifiers to achieve high-level classifier for impetrate to higher recognition rate. This paper proposed method based on Stack Generalization that named Modified Stack Generalization. In rour proposed model, unlike the conventional stacked generalization, the combiner receives the output of base classifiers and original input directly. The experiments have been done on 780 samples of 30 city names of Iran that for different experiments different number of training and testing samples was chosen. In the feature extraction Stage Gradient, Zoning methods are used, and also other method base on Gradient is suggested. Results show that Modified Stack generalization method with the recommended feature extraction method has been achieved to 92.21% recognition rate. Furthermore, Comparison test with other combination methods indicates that the proposed method yields improved recognition rate in the Farsi handwritten word recognition.

Index Terms: Neural Network, Classifier Combination, Classification, Multiple classifier systems

1. Introduction

Handwritten word recognition has attracted a huge scientific interest due to its practicality. Two most important stages in recognition field are feature extraction and classification.

In the feature extraction stage best features must be selected concerning to the recognition domain. There are many feature extraction methods causing the recognition rate's improvement. Two of best feature extraction methods for Farsi word recognition domain are Zoning [1, 2] and Gradient [3].The selection of high performing and scale invariant feature extraction method is an important but difficult task in developing Handwritten Word Recognition. In this paper, we used a feature extraction method, which is describing in section 2, that select scale invariant feature.

There are various methods in classification stage such Bayes-normal [4], Decision Trees [5] Neural Networks [6], Statistical Classifiers [7] and Support Vector, Classifiers [8, 9]. It has become clear that for more complicated data sets like handwritten the traditional set of classifiers would not be suitable to achieve high rate recognition. The recognition rate, however; can be improved by using various types of combining rules. Multiple expert (classifier) decision combination strategies can have a far more reliable and efficient performance in comparison with single expert classifiers.

Instead of looking for the best single classifier, now we look for the best set of classifiers, better to say we seek for the best combination method. Just Imagine; soon we would be able to look for the best set of combination methods, and using all of them depending on the problem complexity [10].

Received: December 16th, 2010. Accepted: June 23rd, 2011

For obtain divers classifiers they should be trained differently [11, 12]. There are three different ways to achieve to this purpose:

1. Different representation of patterns
2. Different learning machines
3. Partitioning the training set
4. Different labeling in learning

There are two main strategies for word recognition: 1- Analytical, 2- Holistic [13, 14]. In first procedure image of the word must be decomposed into sub words or characters by segmentation and in second approach recognition process should be carried out the whole shape of the word Choosing either of the above mentioned approaches is quiet important as it will specify the other stages of the recognition process that we used holistic approach.

There are two main strategies in combining classifiers: fusion (static structures) and selection (dynamic structures) [15]. In classifier fusion, it is supposed that each ensemble member is trained on the whole feature space [16, 17] whereas in classifier selection, each member is assigned to learn a part of the feature space [18- 20]. This way, in the former strategy, the final decision is made considering the decisions of all members, while in the latter strategy, the final decision is made by aggregating the decisions of one or a few of experts.

As described two of the most important procedures in combination domain is static approaches [21-23] and dynamic approaches that static approaches decompose into two main category class conscious and class indifferent [13]. The most important differences between this two domains is, in their combining sector that in this class of committee machines, combiner has to do the act of combining without seeing input data but in the second approach, dynamic structure, the input signal is directly involved in actuating the mechanism that integrates the outputs of the individual experts into an overall output. Average, Min, Max, Product and weighted Ave [24] , Genetic [25,26], Stack Generalization [39] are belongs to the static methods that these techniques for multiple classifier decision combination have been reported extensively in a multitude of task domains which include various text and document analysis problems and cover isolated and cursive handwritten , and printed character or word recognition e.g., Ho et al. [27], Xu et al. [28], Suen et al. [29], Fairhurst and Rahman [30, 31], Yuan et al. [32], Yaeger et al. [33].

Another categorization of static combination strategies is: Non trainable and trainable methods. Genetic Algorithm which is in the static category is application of combining multiple neural networks, such as Cho [34], Yuanhui et al. [35], and Lee [36]. The reconfiguration process customarily involves cycling through the process of training and re-evaluation on the available algorithms, and the novelty of the proposed approach is to streamline and make effective this operation through a process of structured optimization by applying a genetic algorithm [37]. Actually many or even most of the real engineering problems actually do have multiple objectives, i.e., minimize cost, maximize performance, maximize reliability, etc. These are difficult but realistic problems. Genetic Algorithm is a popular meta-heuristic that is particularly well-suited for this class of problems. The concept of Genetic Algorithm was developed by Holland and his colleagues in the 1960s and 1970s [38].

In the static approach it is expected to trainable methods have more accurate results and our experiment indicated that Stack generalization method which is in this category out performs other static methods. Stacked generalization is one of the most popular ways to combining multiple models that have been learned for a classification task [39, 40]. This method is a technique whose purpose is to achieve a generalization accuracy (as opposed to learning accuracy) which is as high as possible. By creating a partition of the learning set, training on one part of the partition, and then observing behavior on the other part, one can try to deduce (and correct for) the biases of one or more generalizers with respect to that learning set [40]. Stack generalization is a mechanism for minimizing the error rate of one or more experts, that combine all of the experts rather than choosing the best one. There are many different ways to implement stacked generalization. Its primary implementation is as a technique for combining

generalizer, although it can also be used when one has only a single generalizer, as a technique to improve that single generalizer [40].

This paper we proposed a method based on Stack Generalization method that named Modified Stack Generalization. Generally the concept of Modified Stack is very similar to stacked generalization but the only difference between Stack and Modified Stack Generalization is in Modified Stack input of the combiner is both output of each expert, and original input. We used extracted features as an input of classifiers in our proposed combination method then the result of this method will be compared with other static methods.

This paper will address the following issues as well. Section 2 discussed about Correlation Reduction strategies. Section 3 describes feature extraction methods and our proposed feature extraction method on the handwritten Farsi word images. Section 4 describes various methods for combining classifier experts. Sections 5 describe Stack and modified Stack generalization. Section 6 experimental results comparing genetic weight optimization, Stack Generalization and Modified Stack Generalization with other combination schemes by using proposed feature are presented. Finally some conclusion remarks are presented in Section 7.

2. Correlation Reduction strategies in multiple classifier systems

The diagram in Figure 1 illustrates four approaches aiming at building ensembles of diverse classifiers. For create diversity between classifiers various methods are exits [41].

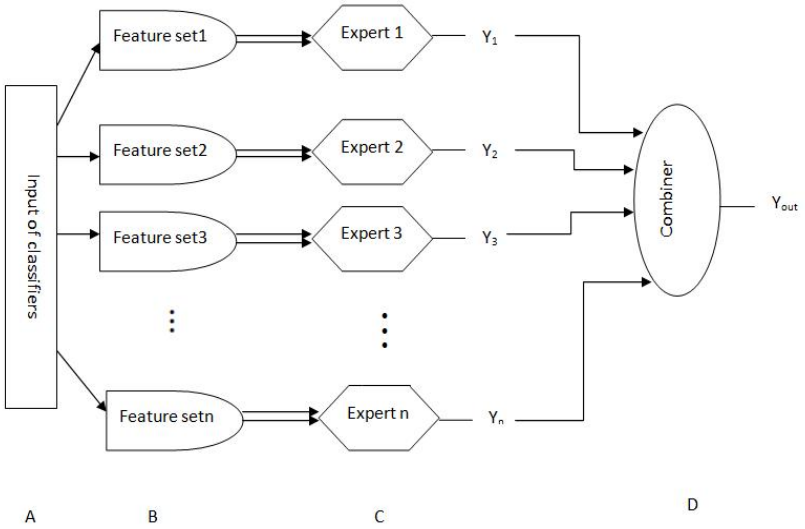


Figure 1. Shows different stage in combination method. Structure of combining methods has four levels. Level A is Data level that use different data subsets. Level B is Feature level that in this level different feature sets extract from input data. Level C is Classifier level that in these level basis classifiers existent and finally Level D is Combination level that in this level combiner must design.

Since classifiers are made through a training procedure, to have classifiers which generalize diversely, they should be trained differently. The training procedure can be affected by input representation of patterns, training samples, learning procedure and supervision strategy, on which correlation reduction techniques will be based on these items, different strategies to make diverse classifiers exist:

A. Different Representation of Patterns

There are three ways to perform this task:

1. Using different feature sets or rule sets.
   This method is useful particularly when more than one source of information is available [38, 41].
2. Using decimation techniques
   Even when only one feature set exists, we can produce different feature sets [42] to be used in training the different classifiers by removing different parts of this set.
3. Feature set partitioning in a multi-net system
   This method can be useful if patterns include parts which are independent.
   For example different parts of identification form or a different handwritten words or digits. In this case, patterns can be divided into sub-patterns each one can be used to train a classifier. One of theoretical properties of neural networks is the fact that they do not need special feature extraction for classification. Feature sets can be the same real valued measurements (like gray levels in image processing). The number of these measurements can be high so if we apply all of them to a single network, the curse of dimensionality will occur. To avoid this problem they can be divided into parts each one used as the input of a sub-network [43, 44] which are independent.

B. Different Learning Machines

There are some free parameters in any learning machine which should be set during training. The final set of these parameters depend on the training set, so even for a given structure and an identical representation of patterns, different training sets could cause different generalizations. Using identical representation of patterns has the advantage that the decision boundary of individual classifiers are in the same axis set (space). Therefore the effect of each sample or expert in composite classifier can be investigated. In these circumstances correlation reduction is based on partitioning the main training set into a few subsets, and using these partitions to train different experts. If different partitions are separated (non-overlapped), the independence of classifiers will be increased, but in most practical cases because of limited number of training samples, these partitions made by perturbing the original training set, have some overlap [46].

C. Different Labeling in Learning

As mentioned earlier, in classification a given pattern x should be assigned to one of several possible classes. Based on this fact, in supervised learning, pairs of input-desired target are used. So changing desired target can change the environment [47].

3. Feature Extraction Methods

The performance of character recognition largely depends on the feature extraction approach. For extract features from word's image various approaches are proposed [48]. That in this section we describe two of better feature extraction methods for Farsi word's image and one new feature extraction method and then in section 3.3 the used feature extraction method in this study is described in details.

A. Gradient

For this type of feature Sobel filters are used. These are two types of filters that detect horizontal and vertical edges .In addition to the values of the horizontal and the vertical Sobel filter, also the absolute value of the filtered image is use to derive the features.

The gradient vector is decomposed into the eight Freeman directions by projecting the vector into the nearest two Freeman direction. Figure2 shows the decomposing the gradient vectors into the eight freeman directions each corresponding to one of the Freeman directions [49, 50].

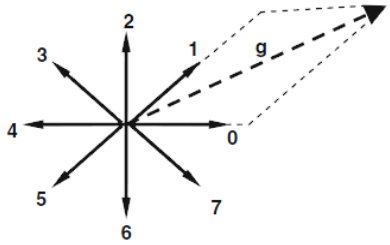


Figure 2. Projecting the gradient vector into the two nearest Freeman directions

B. Zoning

This Feature's are extracted from the normalized character matrix of image [1, 2]. The normalized image is divided into number of zones and in each zone, density of object pixels are calculated. Density is calculated by finding the number of object pixels in each zone and dividing it by total number of pixels having value Black as object pixels [51, 52].

C. Gradient Base Feature Extraction Method

In this section we applied a scale invariant gradient based method for Persian handwritten word recognition feature extraction. It should be note that for this method first, thinning must be applied on the word images. Thinning is the process of reducing thickness of samples to just a single pixel. By the thinning method shape information of patterns preserve and data size reduce [11]. Thinning method removes pixels so that a pattern without holes shrinks to a minimally connected stroke, and a pattern with holes shrinks to a connected ring halfway between each hole and the outer boundary. There, we have used morphology based thinning algorithm for better symbol representation.

After applying the thinning method, the word images decompose into a number of separate images corresponding to four 3×3 masks. Indeed these masks use for scanning lines from 0 to 180 degrees in order to produce their equivalent images. Figure 1 shows these masks. For example the first mask used for separating the lines with 0 degree from the word images. These lines can be extracted as explained below. Instead of input word image considered a zero matrix with the size of original image. This mask moved from left to right over the image and specifies the new values of elements of this matrix. The assigned value of this matrix depends on the value of the pixels in original word image. Scilicet at position of (i,j), the pixels of (i,j), (i,j+1) and (i,j-1) from matrix have 1 value if all of these three pixels from sightly image have 1 value and 0 value, otherwise.

This procedure repeat for other three masks and each word image decompose into four separate images corresponding to these masks. In the other word, with these masks lines with 0, 45, 90 and 135 degrees in word images are separate.

In the next stage, each of separated images uniformly partition into 8 sectors around the center of image. The number of black pixels in each sector is calculating. These values normalize by dividing them upon the total number of black pixels in word images and use for feature value of that sector. Hence for each of these separated images we have eight feature values leading to a feature vector of 32 elements. Thus, we acquire needful feature vector for classification stage. Figure 3 the masks that used for decomposing word images figure 4 shows the stages of this method.

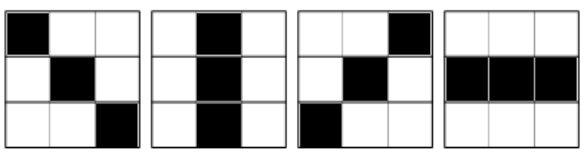


Figure 3. The masks that used for decomposing word images into a number of separate images

Further to the experiments carried out whit this method the results were compared whit two of the best methods in handwritten recognition firm.

4. Combining Methodologies

Assuming $D = \{D1, D2, \cdots, DL\}$ is our L classifiers and $\Omega = \{\omega 1, \omega 2 \cdots, \omega C\}$ is our C class set. Each classifier by receiving $x = R^n$ feature vector, report $D_i(x) = [d_{i1}(x), d_{i2}(x), \cdots, d_{iC}(x)]$ as an output result, that $d_{ij}(x)$ is the belonging precedence of input pattern x to the ω_j. Combining the output of classifiers significance, nomination vector belonging of pattern x to the different classes founded of verdict of L different classifiers. Output of classifiers can organized in the matrix whit Decision Profile titled.

$$DP(x) = \begin{bmatrix} d_{11}(x) & d_{12}(x) & \cdots & d_{1C}(x) \\ d_{21}(x) & d_{22}(x) & \cdots & d_{2C}(x) \\ \vdots & \vdots & \vdots & \vdots \\ d_{L1}(x) & d_{L2}(x) & \cdots & d_{LC}(x) \end{bmatrix} \quad (1)$$

In this matrix i'th row is the output of the Di'th classifier and j'th column is the belonging vector of pattern x to the ω_j 'th class.

There are two general approaches to use DP(x) to find the overall support for each class and subsequently label the input x in the class with the largest support.

- Some methods calculate the support for class i ($(\mu_D^i(x))$) using only the ith column of DP(x). Such methods that use the DP class-by-class will be called class-conscious methods.
  Examples of class-conscious fusion operators are: average, sum, minimum, maximum, product, fuzzy integral, etc.
  The choice of an aggregation method F depends on the explanation of di,j (x), i=1,..,L , j=1,.., c and also is related to characteristic of data.
- Another fusion method is to use all of DP(x) to calculate the support for each class. Fusion methods in this category will be called class-indifferent. Here we can use any classifier with decision profile matrices, as inputs and the class label Dens (x) as the output. There are however some class-indifferent fusion strategies such as decision templates method.

Notice the difference between the class-conscious and class-indifferent groups of methods. The former use the context of the DP but disregard part of the information, using only one column per class, but in the latter methods use the whole DP but neglect the context.

Not that in blew equations x_R is the input pattern and ωC is the arbitrary class.

MIN

The Minimum rule selects by DP matrix, the classifier having the least objection [55].

$$\underset{R=1}{\overset{n}{Min}} rule(x_R) = \underset{k=1}{\overset{C}{MAX}} \left(\underset{i=1}{\overset{L}{MIN}} DP(x_R) \right) \quad (2)$$

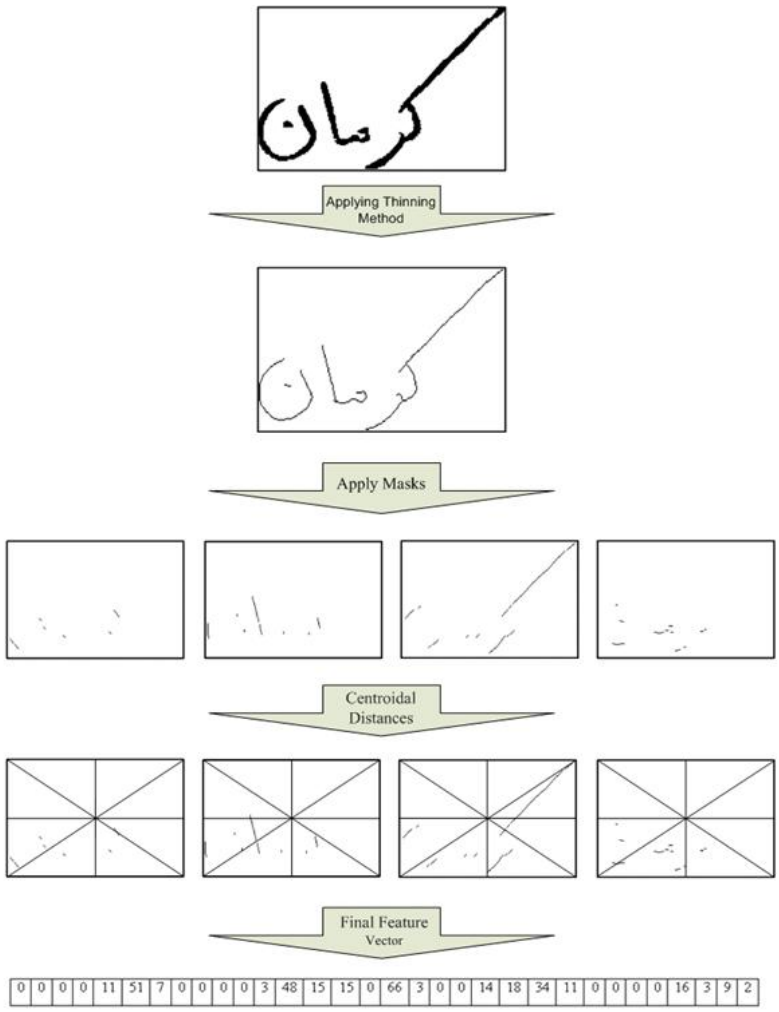


Figure 4. The stages of new feature extraction method for Persian handwritten word recognition. In the first stage the thinning method is applied. In the second stage the word images decompose into a number of separate images corresponding to four 3×3 masks. In the third stage, each of separated images is uniformly partitioned into 8 sectors. Finally, for each of these separated images we have eight feature values leading to a feature vector of 32 elements.

MAX

The Maximum rule in contrast Min rule, selects the classifier producing the highest estimated confidence from DP matrix, which seems to be noise sensitive [55].

$$\underset{R=1}{\overset{n}{Max}}\, rule(x_R) = \underset{k=1}{\overset{C}{MAX}}\left(\underset{i=1}{\overset{L}{MAX}}\, DP\left(x_R\right)\right) \tag{3}$$

PRODUCT

As already pointed out, $P\left(_{x1},\cdots,_{xR}\middle|_{\omega C}\right)$ represents the joint probability distribution of the measurements extracted by the classifiers. Let us assume that the representations used are

conditionally statistically independent. The use of different representations may be a probable cause of such independence in special cases. We will investigate the consequences of this assumption and write.

$$\underset{R=1}{\overset{n}{\Pr o}}\, rule\left(x_R\right) = \underset{k=1}{\overset{C}{MAX}}\left(\prod_{i=1}^{L} DP\left(x_R\right)\right) \tag{4}$$

In product rule (pro), supports provided by the classifiers are multiplied [55].

SUM

In this method the support for ω_j is obtained as the average of all classifiers j'th outputs [55].

$$\underset{R=1}{\overset{n}{Ave}}\, rule(x_R) = \underset{k=1}{\overset{C}{MAX}}\left(\underset{i=1}{\overset{L}{AVR}}\, DP\left(x_R\right)\right) \tag{5}$$

AVERAGE

In this method the support for ω_j is obtained as the average of all classifiers j'th outputs [55].

$$\underset{R=1}{\overset{n}{Ave}}\, rule(x_R) = \underset{k=1}{\overset{C}{MAX}}\left(\underset{i=1}{\overset{L}{AVR}}\, DP\left(x_R\right)\right) \tag{6}$$

For more perception we decided to explain this part with a simple example; assuming we have 5 classifiers (L=5) and 3 classes (c=3). If the result of Decision Profile matrix is as below:

$$DP(x) = \begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0 & 0 & 1 \\ 0.4 & 0.3 & 0.4 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.8 & 0.2 \end{bmatrix} \tag{7}$$

By using explained methods the result of belonging input pattern x to the three classes are obtained as below:

Minimum Rule: $\mu_D(x) = [0 \quad 0 \quad 0.1]$ (8)

Maximum Rule: $\mu_D(x) = [0.4 \quad 0.8 \quad 1]$ (9)

Average Rule: $\mu_D(x) = [0.16 \quad 0.46 \quad 0.42]$ (10)

Product Rule: $\mu_D(x) = [0 \quad 0 \quad 0.0032]$ (11)

Sum Rule: $\mu_D(x) = [o.8 \quad 2.3 \quad 2.1]$ (12)

WEIGHTED AVERAGING

The results from the above illustration should not be taken as evidence that the mean combiner is always the best. The shape of the curve will depend on the problem and on the used base classifier. The average and the product are the two most intensively studied combiners. Yet, there is no guideline as to which one is better for a specific problem. The current understanding is that the average, in general, might be less accurate than the product for some problems but is the more stable of the two [54-58]. But for increasing the result of Average more than Product we can give weight to outputs of experts. Various weighted average combiners have been proposed in the literature.

A. Weights Equal Performance Orrecognition Rates of Each Expert

In Weighed Averaging (WA) model there is one weight per classifier. The support for class ω_j is calculated as:

$$\mu_j(x) = \sum_{i=1}^{L} \omega i \, d ij(x) \tag{13}$$

The weight for classifier d_i is usually based on its estimated error rate [59].

B. Weights Search by Genetic Algorithm

Genetic algorithms (GA) offer a guided random search in the space of all possible feature subsets [51]. Using genetic algorithms for classifier selection pays off when the search space is large and the calculation of the criterion function (ensemble accuracy, ensemble diversity, individual accuracy, and so on) is cheap. Genetic algorithms are employed to evolve these weights so that they can characterize to some extent the fitness of the classifiers to join the ensemble [51]. At present, there are many different methods to recognize the Farsi words. The use of genetic algorithm to recognize a character has been a new algorithm used in this problem. Genetic algorithms offer a particularly attractive approach for this kind of problems since they are generally quite effective for rapid global search. Moreover, genetic algorithms are very effective in solving large-scale problems.

Genetic Algorithm (GAS) is a search technique used in computer science to find approximate solutions to optimization and search problems and is inspired by evolutionary biology such as inheritance, mutation, natural selection and recombination. Genetic algorithms are typically implemented as a computer simulation in which a population of abstract representations of candidate solutions to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but different encodings are also possible. The evolution starts from a population of completely random individuals and happens in generations. In each generation, the fitness of the whole population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), modified (mutated or recombined) to form a new population, which becomes current in the next iteration of the algorithm [25-26].

TRAINABLE AND NON-TRAINABLE (STATIC)

To perfect analyzing the various combining structure we investigate learning type of classifier which is used as a combiner. The combining classifiers might be defined as a trainable combiner or non-trainable combiner. In non-trainable classifiers category, combiners do not need training after the classifiers in the ensemble have been trained individually. Two examples of this group are averaging and DTs methods (Decision Templates). Other combiners

need additional training before or during training of the individual classifiers, for example, the weighted average combiner, SG (Stacked Generalization), AdaBoost and Mixture of Experts. Stack Generalization is one of the most famous methods which is described in subsection 4.1 and our proposed method which is essentially based on stack generalization is depicted in details in subsection 4. 2

STACK GENERALIZATION

Stacked generalization is a technique proposed by Wolpert [40] that extends voting in those learners (called level 0 generalizers) are not necessarily combined linearly [60] base classifiers stand in this level . The combination is made by a combiner gate (called level 1 generalizer) that is also trained by out- put of level 0. So general framework of this method consists of two levels, fig3 shows these levels. First, the level-0 networks are trained, using the input data and the target outputs. Then the outputs of the first layer with the corresponding target class are used to train the level-1 network. At the end final decision has been picked up by the level 1[61]. Figure5 shows the structure of Stack Generalization method.

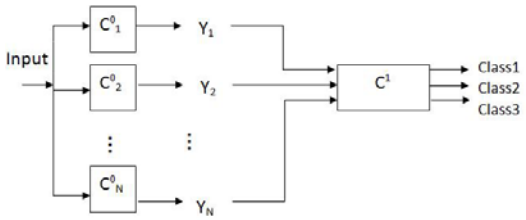


Figure 5. Block diagram of a multiple neural networks system based on stacked generalization

MODIFIED STACKED GENERALIZATION METHOD

Generally the concept of modify stack is very similar to stacked generalization, unless that the level-1 classifier at the training phase, in addition of level 0's outputs, trained by training dataset from the original input. This is caused to make new matrix for mapping the input of level 1. In fact, this method tries to find a weighting system which will specify the best combination of the base classifiers. Block diagram of this method is shown in fig 6.

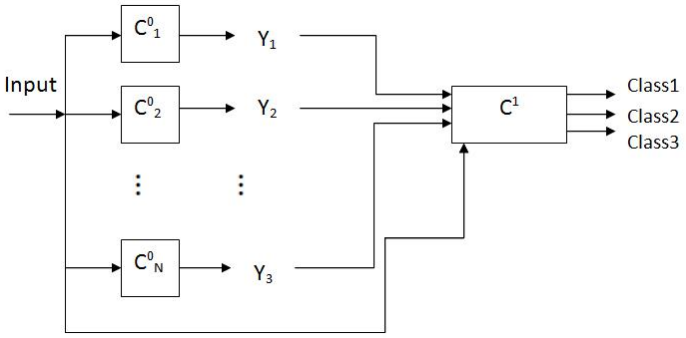


Figure 6. Block diagram of modified stacked generalization

5. Experimental Results

The discrepancy between different combining methods (discussed in section 4) has been well tested via several experiments using a dataset of Iran cities names. The dataset used in this paper consist of 780 samples of 30 cities names for each city name 26 samples will be accessible which are written by 26 different people. In our experiment we categorize entire dataset into three sets. The first test set includes, 17 samples being selected for training and 9 for test, the training set will consequently consist of 510 samples and test set will include 270 samples. In the other set we increased the number of train samples to 19 and decreased train samples to 7 so training set will consist of 570 samples and test will include 210 samples. For the third experiment shuffled data are used. In the third set we used shuffled data and then we select 20 samples from each city name for training and 6 sample for test set so training set consist of 600 samples and test set will include180 samples. All samples will be scanned at 96 dpi resolution in the gray scale format. All samples shall be converted into binary format with a constant threshold value prior to be used at the feature extraction stage. Images were centralized in a 184×320 pixels frame in the last stage. Some sample images are shown in Fig7.

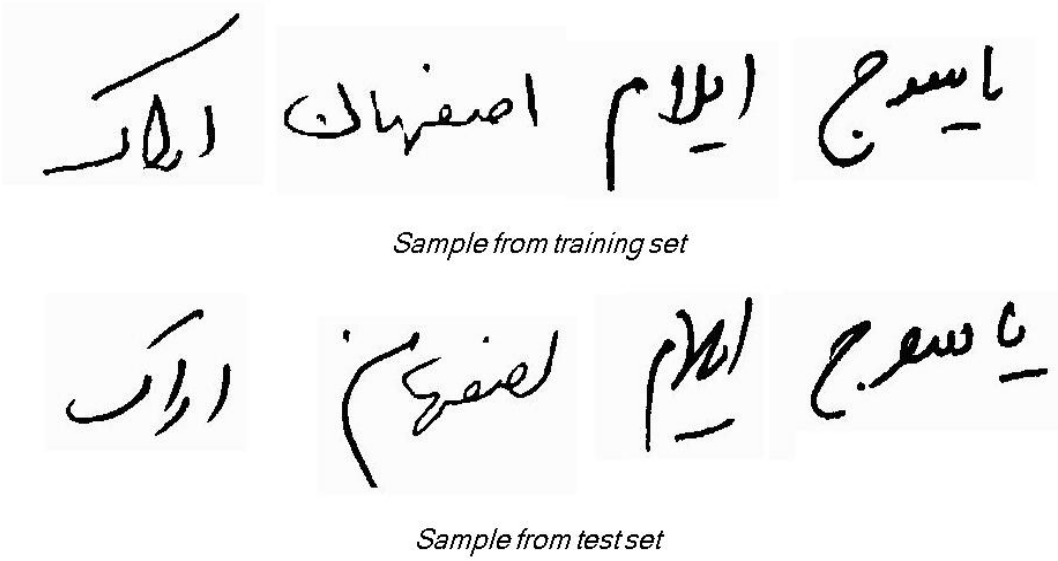
Through various experiments the results of seven combination methods in the Farsi handwritten word recognition domain via multi-layer perceptron (MLP) classifier are comprised.

The MLP has different learning parameters, such as; number of epochs, learning rate and number of neurons in hidden layer. The required number of epochs to reach the highest recognition rate was estimated by four-fold cross validation on the training set. Figure 8 shows the diagram of selected hidden neurons.

A. Experiment 1

In this experiment all three sets are used. Categorized results were compared through different combination methods, outlined in section 4, these results, are summarized in Table 2.

These results were obtained through testing nine different combination methods using all three dataset with multi-layer perceptron (MLP) as basic experts. Toward creating diversity between basis experts we used different parameters for each expert which is shown in Table 1.



Sample from training set

Sample from test set

Figure 7. Samples of city names of Iran from training and test set

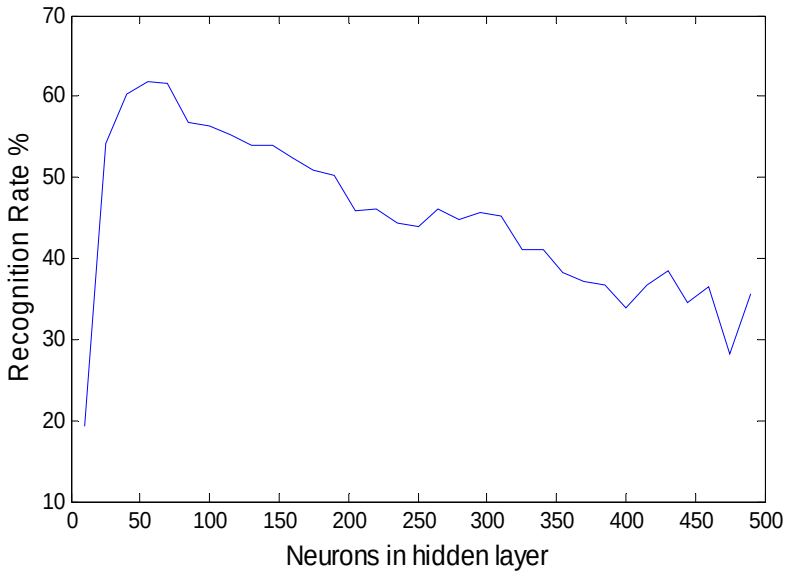


Figure 8. Diagram of selected hidden neurons

Table 1. Specification for each expert

	No. of hidden neurons	**Eta**	**Iteration**	**Best Recognition rate**
First expert	45	0.15	400	73.87
Second expert	47	0.15	400	73.12
Third expert	50	0.1	400	73.9
Forth expert	40	0.1	400	72.60

Confusion matrix is used to show the diversity between experts. Figure9 shows these confusion Matrices. In this experiment the result of one MLP for first set is 70.32. In artificial intelligence, a confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. One of the advantages of the confusion matrix is the easiness in managing the cases that the system confused two classes. When a data set is unbalanced (when the quantity of the samples in different classes have a great variation) error rate of the classifier would not be not representative of the true performance of the classifier [57].

The results for the average accuracy for each combination method in 10 times run are shown in Table 2. In this experiment Gradient base feature extraction method was applied for all three sets and after different combination methods were utilized and Modified Stack generalization method were compared with other static methods, for further details please refer to section 4.

To evaluate the performance of proposed model and also exhibit the advantage of using it in recognition of Farsi word, it is compared with other fusion methods such as Sum ،Min ، Max ،Average, product, and weighted average aggregation rules on Iranshahr dataset. The best result of each method is shown in table 4.

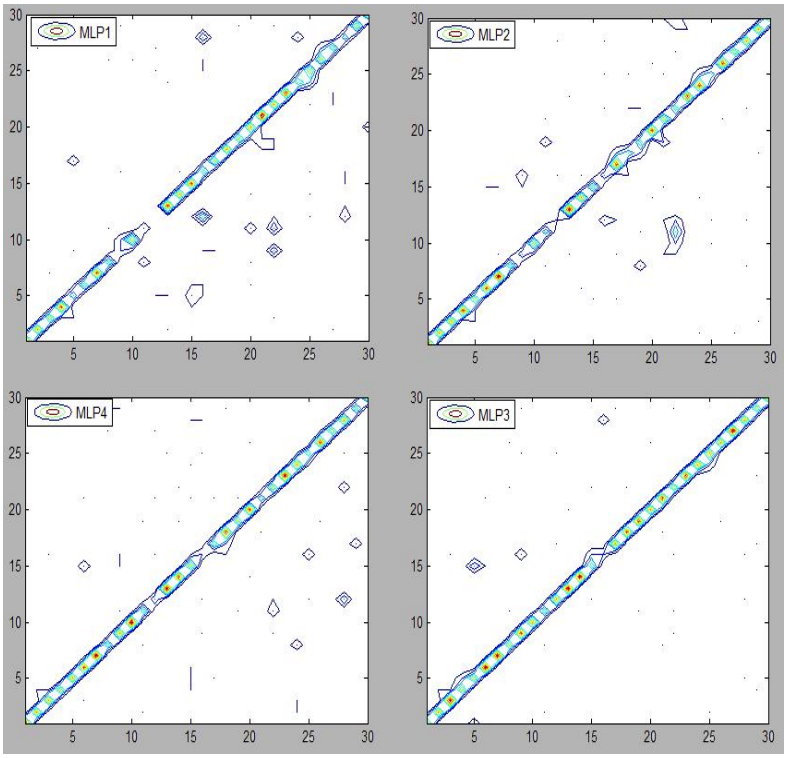


Figure 9. Confusion Matrix of individual classifiers. As you can sea in these pictures the diagonal has the most density and there are few scattered dotes around indicating to the wrong classifications for each expert.

Table 2. Recognition rate of different combination methods on the three test set. Referring to the third row of bellow table, GA method has 0.20% performance improvement with regards to weighted averaging. Modified Stack in comparing with GA (in third set) has 1.39% improvement. The best performances in each row are bolded. As it seen from the row, modified stack method has a 1.24% improvement in recognition rate.

As expected and described before between combining methods Stack Generalization and Modified Stack are two of good algorithms and our proposed method in comparison between other described class conscious methods has a great improvement in recognition rate. Also as it represented in table 3, comparison between Stack Generalization and Modified Stack (for third set) has 81.32% and 92.21% increment. This improvement is because of the combiner receives the input pattern directly adding on the base classifiers outputs.

It is exhibited in Table 3 that in the static methods weighted averaging, genetic algorithm and stack generalization are the most successful method. With more accurate outlook to Table 3 in comparison with various class conscious methods, Weighted Averaging will have the second ranking after Genetic Algorithm. Weighted Averaging is one of the good methods.

Static Approches									Dynamic Approch
Fusion methods	**Min**	**Max**	**Product**	**Sum**	**Averaging**	**Weighted Averaging**	**Genetic Algorithm**	**Stack Generalization**	**Modified Stack**
First Set	79.04	76.44	83.33	83.07	82.52	85.74	86.68	86.9	**87.11**
Second Set	81.88	83.22	86.66	85.44	86.33	87.23	88.16	89.02	**90.09**
Third Set	82.94	85.22	88.5	87.38	88.22	88.77	88.97	89.12	**90.36**

Table 3. The best result of different combination methods. Referring to the third row of bellow table, GA method has 1.23% performance improvement with regards to one of the best method in recognition domain weighted averaging methed. Modified Stack in comparing with GA(in third set) has 1.11% improvement. The best performances in each row are bolded.

Static Approches									Dynamic Approch
Fusion methods	**Min**	**Max**	**Product**	**Sum**	**Averaging**	**Weighted Averaging**	**Genetic Algorithm**	**Stack Generalization**	**Modified Stack**
First Set	81.67	77.9	85.12	84.56	84.5	86.65	87	87.89	**88.32**
Second Set	83	85.12	87.68	86	87.12	88.38	90	90.24	**91.9**
Third Set	85.4	87	89.32	88.9	89.45	89.87	91.10	91.32	**92.21**

The difference between first set and second set is due to the number of the training sets and as we know number of the train sets has an important impact on this issue. As a case in point referring to the table3 **,** GA method has 1.23% performance improvement with regards to one of the good method in recognition domain ,weighted averaging method and also our proposed method has 3.89% and 0.31 Improvement in recognition rate with respect to first set and second set respectively.

Since the first set and second set were not shuffled, it was possible to have hard samples in test sets and in training phase, by using shuffled data we can solve this problem.

Referring to table 2 in comparison between Stack Generalization and Modified Stack, for third set, by using Modified Stack the recognition rate has 1.24% increase. This improvement is because of the combiner receives the input pattern directly adding on the base classifiers outputs. By comparison between last column in table1 and all recognition rates in both table2 and 3 it is become obvious that the recognition rate of combining methods are higher than using one MLP.

B. Experiment II

Referring to the previous experiment the Modified Stack method with high variation was more accurate in comparison to the other methods, as well as, the applied feature extraction method which is more suitable comparing to the other methods.

To show the advantage of our applied feature extraction method, the result of combining methods by using this method with two famous feature extractions in the handwritten recognition domain (Zoning and gradient) is compared. Such as previous experiment for combination of classifiers, 4 MLP as basic classifiers was used. The third set was used for this experiment. The results of this experiment are tabulated as bellow (Table4).

Table 4. Different feature extraction methods are applied to different combining methods. The best result in each column is underlined

	Static Approches								Dynamic Approch
Fusion methods	**Min**	**Max**	**Product**	**Sum**	**Averaging**	**Weighted Averaging**	**Genetic Algorithm**	**Stack Generalization**	**Modified Stack**
Zoning	73.15	75	76.04	74.8	79.22	80.98	83.5	86.11	**87**
Gradient	75.78	79.22	81.1	68.33	80.34	83.22	87.78	87.9	**88.3**
Gradient base	82.94	85.22	88.5	87.38	88.22	88.77	88.97	89.12	**90.36**

As you can see in comparison between different feature extraction methods, gradient base feature extraction method is the best one and Gradient is in the second rank. As it described in section III.3, the Gradient base method has more concerning process. So as it obvious in the table4 in all of combining methods this feature extraction method has a perceivable improvement in recognition rate. As it mentioned in the later experiment our proposed method, Modified Stack Generalization, has a good performance. In this section with respect to this method the zoning method has 87% recognition rate, and gradient method has 88.3% recognition rate whereas our utilized method has 90.36% recognition rate. This improvement is observable in all of combining methods. In other words in each column of table 4 the noticeable improvement is mentioned. Thus with respect to both applied combining method and feature extraction method this study is optimal in the matter of complexity of Persian handwritten word problem.

6. Conclusions

The recognition results that were reported in this paper show the comparison between different combining methods in Farsi Hand Written Words on Iranshahr datasets. Considering different combining methods, the recognition rates were 82.94%, 85.22 %, 88.5%, 87.38%, 88.22%, 88.77%, 88.97%, 89.12% and 90.36% for Min, Max , Product, Sum, Average, Weighted Average, Genetic Algorithm, Stack Generalization And Modified Stack respectively. As you can see Modified Stack Generalization has the highest recognition rate because in this method classifier in the gating part (level-1) classifier at the training phase, in addition of level 0's outputs, trained by training dataset from the original input.

References

[1] L. Rokach, "Ensemble-based classifiers," *Artif Intelligent Rev*, Vol. 33, pp.1–39, 2010.

[2] H. Aljuaid, Z. Muhammad and et al, "A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach," *Journal of Computer Science* No. 6, Vol. 6, pp. 619-624, 2010.

[3] N. Ahmad, T. Natarjan, and K,Roa, "Discrete Cosine Transform," *IEEE. Trans Compute*., Vol. C-23, pp. 90-93, 1974.

[4] H. Choi, S. J. Cho, and J. H. Kim "Generation of Handwritten Characters with Bayesian network based On-line Handwriting Recognizers," *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)* 0-7695-1960-1/03 $17.00 © 2003 IEEE

[5] F.ch Li and F. Guan: 'heuristic Model Research on Decision Tree Algorithm'. Intelligent Interaction and Affective Computing, 2009. ASIA '09. *International Asia Symposium on* pp. 149 – 152, 2009.
[6] DS. Lee, and SN. Srihari, "A theory of classifier combination: the neural network approach." *Proceedings of the third international conference on document analysis and recognition, Montreal, Canada*, pp. 42–5, 1995.
[7] G. Giacinto, F. Roli, and L, Bruzzone, "Combination of neural and statistical algorithms for supervised classification of remote-sensing images," *Pattern Recognition Letter*s, Vol. 21, pp. 385-397, 2000.
[8] CJC. Burges, "A tutorial on support vector machines for pattern recognition," *Knowl Disc Data Min*, Vol. 2, pp.1–43, 1998.
[9] G. M. FUNG, "Multi category Proximal Support Vector Machine Classifiers," *Machine Learning*, Vol. 59, pp. 77–97, 2005.
[10] T. K. Ho, "Multiple classifier combination: Lessons and the next steps. In A. Kandel and H. Bunke editors, Hybrid Methods," *Pattern Recognition. World Scientific Publishing*, pp. 171–198, 2002.
[11] L. I. Kuncheva, "Combining Pattern Classifiers, Methods and Algorithms," *INC., PUBLICATION*, 2004.
[12] R. Polikar, "Ensemble Based sysems in Decision Making". *IEEE CIRCUITS AND SYSTEMS MAGAZINE*, Vol. 06, pp. 1531-1636, 2006.
[13] S. J. Soltysiak, "Visual information in word recognition: word shape or letter identities," in: *Proceeding Workshop Integration of Natural Language and Vision Processing*, pp. , 1994.
[14] M. Leung, and A. M. Peterson, "Scale and Rotation Invariant Texture Classification," *Proceeding International Conference Acoustics, Speech, and Signal Processing*, pp. 461–165, 1991.
[15] K. Woods, W. P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," IEEE Trans. Pattern Anal. Mach. Intell, Vol.19 , pp.405-410, 1997.
[16] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their application to handwriting recognition," *IEEE Trans. Systems Man Cybernet*. Vol.22, pp. 418-435, 1992.
[17] K. -C. Ng, B. Abramson, "Consensus diagnosis: a simulation study," *IEEE Trans. Systems Man Cybernet*. Vol.22, pp. 916-928, 1992.
[18] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput*. Vol.3, pp. 79-87, 1991.
[19] L. A. Rastrigin, and R. H. Erenstein, "Method of Collective Recognition," *Energoizdat, Moscow*, 1982 .
[20] Alpaydin, and M. I. Jordan, "Local linear perceptrons for classification," *IEEE Trans. Neural Networks*, Vol.7, pp. 788-792, 1996.
[21] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their application to handwriting recognition,' *IEEE Trans. Systems Man Cybernet*. Vol.22, pp. 418-435, 1992.
[22] K.-C. Ng, and B. Abramson, "Consensus diagnosis: a simulation study,' *IEEE Trans. Systems Man Cybernet*. Vol.22, pp. 916-928, 1992.
[23] L. I. Kuncheva, James C. Bezdek, and Robert P.W. Duin, "Decision Templates for Multiple Classifier Fusion: *An Experimental Comparison," Pattern Recognition*, vol. 34, pp. 299-314, 2001.
[24] C. A. Shipp, and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Inform Fusion*, Vol. 3, pp.135–48, 2002.
[25] J. Kennedy, RC. Eberhart, "Particle swarm optimization," *Proceedings of the IEEE international conference on neural networks (ICNN)*, Vol.4, pp. 1942–1948, 1995.

[26] R. KALA, H. VAZIRANI, A. SHUKLA and R. TIWARI, "Offline Handwriting Recognition using Genetic Algorithm," *IJCSI International Journal of Computer Science Issues*, Vol. 7, Issue 2, No 1, March 2010.
[27] T. K. Ho, J. J. Hull, and S.N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans Pattern Anal Intell*, Vol. 16, pp. 66–75 Mach 1994.
[28] L. Xu, A. Krzyzak, and C.Y. Suen, "Associative switch for combining multiple classifiers," *Joint Conf. on Neural Networks*, vol.1. pp 43–48, 1991.
[29] C. Y. Suen, C. Nadal, T. A. Mai, R. Legault, and L .Lam, "Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts," *Proc. IWFHR, Montreal, Canada*, pp 131–143, ,1990.
[30] M. C. Fairhurst, A. F.R. Rahman, "A generalised approach to the recognition of structurally similar handwritten characters," *IEE Proc Vision Image Signal Process* Vol.1, pp.15–22, 1997.
[31] A. F. R. Rahman, and M. C. Fairhurst, "Enhancing multiple expert decision combination strategies through exploitation of a priori information sources," *IEE Proc Vision Image Signal Process* Vol.1, pp.1–10, 1999.
[32] T. Yuan, T. Lo-Ting, L, Jiming, L. Seong-Whan, and L. Win-Win, "Off-line recognition of Chinese handwriting by multifeature and multilevel classification," *IEEE Trans Pattern Anal Mach Intell*, Vol 5, pp.556–561,1998.
[33] L. S. Yaeger, B. J. Webb, R. F. Lyon, " Combining neural networks and context driven search for online, printed handwriting recognition in the NEWTON," *AI Mag19* Vol.1, pp.73–89 ,1998.
[34] Sung-Bae, "Combining modular neural networks developed by evolutionary algorithm,: *Conf. on Evolutionary Comput*, Vol. , pp.647–650, Indianapolis,Ind., USA,1997.
[35] Z. Yuanhui, Z .Zhaohui, L. Yuchang, and S. Chunyi Multistrategy, "learning using genetic algorithms and neural networks for pattern classification," *on Systems, Man, and Cybernetics: Information Intelligence and Systems* Vol. , pp. 1686–1689, Beijing, China 1996.
[36] L. Seong-Whan, "Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network," *IEEE Trans Pattern Anal Mach* 18(6):648–652,(1996).
[37] L. S. Oliveira And R. Sabourin, "A Methodology For Feature Selection Using Multiobjective Genetic Algorithms For Handwritten Digit String Recognition," *International Journal of Pattern Recognition and Artificial Intelligence* Vol. 17, pp. 903-92, 2003.
[38] JH. Holland, "Adaptation in natural and artificial systems," *University of Michigan Press*; 1975.
[39] K. Ming Ting, and I. H. Witten, "Stacked Generalization: when does it work" *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pp. 866-871, 1997.
[40] H. Wolpert, "Stacked Generalization," *Neural Networks*, Vol. 5, pp. 241-259.
[41] R. ghaderi, "Arranging simple neural networks to solve complex classification problems," *PHD thesis from University of Surrey*, 2000.
[42] J. Kittler, A. Hojjatoleslami, and T. Windeatt, "Weighting factors in multiple expert fusions," *In Proc. of British Machine Vision Conference BMVC97, Essex University, Essex U.K*, pp. 42-50 1997.
[43] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," Connection Science, special issue on combining arti_cial neural ne tworks: *ensemble approaches*, Vol. 8, pp 385-404. 1996.
[44] H. A. Rowley, "Neural Network-Based Face Detection," PhD thesis, School of computer science, *Computer scince Department, Carnegie Mellon universit, Pittsburg, PA* 15213, 1999.

[45] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. on Pattern analysis and Machine Intelligence*, 20,Vol. 1, pp.23-28, 1998.
[46] T. Windeatt and R. Ghaderi, "Dynamic weighting factors for decision combining," *International Conf. on Data Fusion, Great Malven, U.K.*, Vol. pp. 123-130, 1998.
[47] R. Ghaderi, "Arranging simple neural networks to solve complex classification problems," *Submitted for the Degree of Doctor of Philosophy from the University of Surrey*.
[48] Sh. Abdleazeem, and E. EL-sherif, "Arabic handwritten digit recognition," *IJDAR, Vol*.11, pp.127-141, 2008.
[49] N. Ahmad, T. Natarjan, and K,Roa, "Discrete Cosine Transform," *IEEE. Trans Compute*., Vol C-23, PP 90-93, 1974.
[50] C. L. Liu., Y. J. Liu., and R. W. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition," A.C. Downton, S. Impedovo (Eds.), *Progress of Handwriting Recognition, World Scientific, Singapore*, Vol. , pp. 161–168,1997.
[51] G. Fumera and F. Roli, "Performance analysis and comparison of linear combiners for classifier fusion," *In Proc. 16th International Conference on Pattern Recognition, Canada*, Vol. ,pp. 2002.
[52] M. Bokser, "Omni Document Technologies," *Proceedings of the IEEE*, vol.80 .pp 10661077, July 1992.
[53] Sh. Abdleazeem, E. EL-sherif, "Arabic handwritten digit recognition," IJDAR, Vol.11, pp.127-141, 2008.
[54] P. W. Robert, D. David, and M.J. Tax, "Experiments with Classifier Combining Rules," *Pattern Recognition Group, Department of Applied Physics Delft University of Technology, LNCS* 1857, pp. 16−29, 2000.
[55] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactionson Pattern Analysis and Machine Intelligence*, 20 Vol.3, pp.226–239, 1998.
[56] R. P. W. Duin and D. M. J. Tax, "Experiments with classifier combination rules," *In J. Kittler and F. Roli, editors, Multiple Classifier Systems, of Lecture Notes in Computer Science, Cagliari, Italy, Springer*, pp. 16–29.2000.
[57] J. Miller and L. Yan. Critic-driven, "ensemble classification," *IEEE Transactions on Signal Processing*, 47 Vol.10, pp. 2833–2844, 1999.
[58] M. J. Tax, R. P. W. Duin, and M. van Breukelen, "Comparison between product and mean classifier combination rules," *In Proc. Workshop on Statistical Pattern Recognition, Prague, Czech*, 1997.
[59] M. J. Tax, M. van Breukelen, R. P. W. Duin, and J. Kittler, "Combining multiple classifier by averaging or multiplying," *Pattern Recognition*, Vol. 33, pp. 1475–1485, 2000.
[60] D. J. Romero, L. M. Seijas, "Directional Continuous Wavelet Transform Applied to Handwritten Numerals Recognition Using Neural Networks," *JCS&T*, Vol. 7 No. 1, 2007.
[61] Sajedin, Sh. Zakernejad, S. Faridi, M. Javadi and R. Ebrahimpour, "A Trainable Neural Network Ensemble for ECG Beat Classification", *In Proceedings of the International Conference on Neural Networks (ICNN2010), Amsterdam, Netherland*, September 28-30, 2010.

Reza Ebrahimpour is an Assistant Professor at the Department of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. He is also a research scientist at School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran. He received the B.S. degree in Electronics Engineering from Mazandaran University, Mazandaran, Iran and the M.S. degree in Biomedical Engineering from Tarbiat Modares University, Tehran, Iran, in 1999 and 2001, respectively. He obtained Ph.D. degree in the field of Computational Neuroscience from the School of Cognitive Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran in July 2007. Dr. Ebrahimpour is the author or co-author of more than 60 international journal and conference publications in his research areas, which include Human and Machine Vision, Computational Neuroscience, Neural Networks, Pattern Recognition and Multiple Classifier Systems.

Mona Amini was born in Tehran, Iran on August 13, 1984. She received the bachelor degree in Electronics from Azad University, Qazvin in 2007 and now she is student in master of Mechatronics in Azad University of South Tehran. She interests is researching in the areas of neural networks, handwritten recognition specially on combining classifiers and Farsi handwritten.

Fatemeh Sharifizadeh was born in Tehran, Iran. She received B.Sc. degrees in Computer Science from the University of Tehran, in 2009. She is now a student of Computer Science in M.Sc. degree. Her interest areas are artificial intelligence, Machine Learning, Intelligent Systems, Pattern Recognition, and Computational Neuroscience.