Ensemble Convolutional Neural Networks for Face Recognition

Wen-Chang Cheng[†]
Department of Computer Science
and Information Engineering
Chaoyang University of
Technology
Taichung Taiwan
wccheng@cyut.edu.tw

Tin-Yu Wu
Department of Computer Science
and Information Engineering
Chaoyang University of
Technology
Taichung Taiwan
s10727603@gm.cyut.edu.tw

Dai-Wei Li
Department of Computer Science
and Information Engineering
Chaoyang University of
Technology
Taichung Taiwan
s10627616@gm.cyut.edu.tw

ABSTRACT

Many methods about face recognition have been put forward. Generally, the feature of the face image is extracted and then the classifier is used to complete the face recognition. There is no direct relationship between the method of feature extracting and the classifier. In recent years, based on the Convolutional Neural Networks of deep learning, better effect is achieved by merging feature extraction and classifier. What's more, ensemble classifier has become the common method of improving classifier accuracy. Therefore, in this paper, we have proposed two kinds of methods based on multiple CNNs classifiers for face recognition. These method has the same structure of CNNs but with different training sets. The first kind of method, adopting the way of random sampling and replacing, trains individual CNNs by using different training sets produced in the original set of data. The second kind of method produces different training sets by enlarging and decreasing the image size of the original set of data. The two kinds of method eventually use the voting method to merge individual CNNs results. Experiments have proved that the ensemble CNNs classifier is better than the single CNNs classifier, and the accuracy of the face recognition is as high as 99.5%.

CCS CONCEPTS

Computing methodologies→Machine learning→Machine learning algorithms→Ensemble methods; Computing methodologies→ Artificial intelligence→Computer vision→Computer vision problems

KEYWORDS

Deep Learning, Ensemble classifier, Multilayer feature representation, Image classification

ACM Reference format:

Wen-Chang Cheng, Tin-Yu Wu and Dai-Wei Li. 2018. Ensemble Convolutional Neural Networks for Face Recognition. In *Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 18). Sanya, China, 6 pages.* https://doi.org/10.1145/3302425.3302459

1 Introduction

Face recognition is a common research topic in pattern recognition, so many methods have been proposed. These methods are used for feature extraction in face images, and then the classifier is used to complete face recognition. The used facial features include local features, global features, or both, while the common classifiers include neural networks, support vector machines, AdaBoost, etc. Nevertheless, calculation methods of these features are designed on the basis of observation or mathematical model of the expert, and better results are obtained in combination with test classifier, so that there is no direct relationship between the feature and the classifier. The problem has been greatly improved through deep learning [1-5]. Deep learning is a way of extraction and expression of multilayer information features. It also directly merges neural network and good feature extraction and recognition accuracy can be obtained through extensive training of data learning.

Convolutional neural networks (CNNs) is kind of deep learning neural network which simulates the human visual identification system. CNNs has been successfully adopted in many image classification problems [6-13]. CNNs is the combination of convolutional layer, pooling layer and fully-connected network layer. The convolutional layer is commonly used interchangeably with the pooling layer to generate multilayer information features. Also, different number of layers can be used. The fully-connected network layer can form a multilayer perceptron (MLP), which is used to identify the classification of the feature vectors generated by the multilayer convolutional layer and pooling layer previously. Ensemble classifier has become the common method of improving classifier accuracy [14]. It is a decision method by integrating multiple identical or different classifier results, which can make decisions by effectively collecting different opinions of experts, better than the single expert opinion. Also, ensemble classifier can achieve better effect than single classifier. Common ensemble methods of ensemble classifier include bagging, boosting, random forest and so on [9-14]. Therefore, some scholars has put forward

[†]Corresponding author: wccheng@cyut.edu.tw

ACAI '18, December 21–23, 2018, Sanya, China
© 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6625-0/18/12…$15.00
https://doi.org/10.1145/3302425.3302459

that complete image classification by combing many CNNs. These methods generate different new training sets from the original data sets to train individual CNNs, and finally add the individual CNNs's classification probability as the final classification results. These methods are applied to some classified benchmark data sets and the best results are obtained.

In this paper, we have proposed two kinds of methods based on multiple CNNs classifiers for face recognition. The first kind of CNNs ensemble classifier, adopting the way of random sampling and replacing, trains individual CNNs by using different training sets produced in the original set of data and combines individual CNNs results through voting method. The second kind of method produces different training sets by enlarging and decreasing the image size of the original set of data of the CNNs ensemble classifier and also combines individual CNNs results through voting method. It is proved by the experiment of face database [15] that the CNNs ensemble classifier is better than the single CNNs. Of the paper structure, the second section introduces the CNNs structure used in the paper. The third section puts forward two kinds of CNNs ensemble classifiers. The fourth section are experimental results and discussions and the last section is the conclusion.

2 Convolutional Neural Networks

As has been discussed above, CNNs is the combination of convolutional layer, pooling layer and fully-connected layer. The convolutional layer makes the input image go through the fixed size filter to perform the whole image convolution operation, and a filter can obtain a feature map after convolution operation. Different filters will obtain different feature maps and the convolutional layer often uses different filters. The pooling layer performs sampling operations on the fixed size region of the input feature map and common sampling methods include max sampling, mean sampling, random sampling and so on. The fully connected layer compose a Multilayer Perceptron (MLP). The input image is classified by the feature vectors of the multilayer convolution layer and the pool layer.

Figure 1 is an example of a five-layer CNNs [1]. The first layer is convolutional layer, which uses four filters to generate four feature maps of the input images (shown in Figure 1 (S1)). Assume that the input image size is $n \times n \times 1$ pixels (width and height is n, the channel number is 1) and the filter size is $k \times k \times 1$ pixels (width and height is k, the channel number is 1), then the output feature map size is $n \times n \times 4$ pixels (ignoring the convolution edge effect). The second layer is the pooling layer, which generates four new feature maps through area sampling of the input four feature maps (shown in Figure 1 (C1)). Assume that the sampling area is 2×2 pixels without overlapping, then the size of the output feature map is $(n/2) \times (n/2) \times 4$ pixels. The third layer is the second convolutional layer which uses six filters. Assume the pixel of each filter size is $k \times k \times 4$, six new feature maps are generated by convolution operation of the four feature maps with the same position (shown in Figure 1 (S2)). Then the output feature map size is $(n/2) \times (n/2) \times 6$ pixels (ignoring the convolution edge effect). The fourth layer is the second pooling layer, which generates six feature maps by area sampling (2×2 pixels) the six feature maps of the last layer (shown in Figure 1 (C2)) and the output feature map size is $(n/4) \times (n/4) \times 6$ pixels. The fifth layer is a fully connected two layer perceptron (shown in Figure 1-the part of fully connected MLP). The $(n/4) \times (n/4) \times 6$ pixels are rearranged into a feature vector and input to the perceptron, and finally output the results of classification. Each layer output, in addition to the pooling layer, all need to go through an activity function operation, and then output again. The common activity functions include *sigmoid*( ), *tanh*( ) or *Relu*( ). While, the activity functions commonly used by the output layer is *softmax*( ) or linear function.

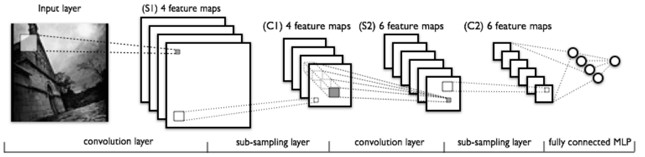

Figure 1: An Example of Five-layer CNNs [1].

In the following section, we will introduce two kinds of ensemble classifiers based on multiple CNNs and the used CNNs structure is the same as in this section.

3 CNNs-based Ensemble Classifier

In this section, we will introduce two kinds of ensemble classifiers based on multiple CNNs. To enable different CNNs in an ensemble classifier to have different classification capabilities, we use different training sets to train individual CNNs. The first kind of ensemble classifier, adopting the way of random sampling and replacing, trains individual CNNs by using different training sets produced in the original set of data. The second kind of ensemble classifier produces different training sets by enlarging and decreasing the image size of the original set of data. The two kinds of method eventually use the voting method to merge individual CNNs results. The two kinds of methods will be explained below.

3.1 Method 1

Figure 2 is the structure of the first ensemble classifier, in which the structure of each CNNs is the same as Figure 1. Each CNNs uses different training sets to train so as to get different CNNs with different identification abilities. The following will introduce them in two aspects: training phase and the test phase. In the training phase, assuming an original face training set, we randomly extract the face image with the same number of original training sets as the new training set, and then randomly extract the face image back into the original face training set. Therefore, a face image may reoccur in the new training set, or may not occur in the new training set, and the same face image in different order of the new training set may be different in order to generate a new training set to train the individual CNNs, aiming to produce different recognition ability of CNNs. In the test phase, an unknown face image is input, and each CNNs is computed to obtain the individual result. Finally, the result is merged by voting method. Scholars such as D. C.

Ciresan [9-11] has put forward that in deep learning ensemble classifier, the aggregation method of adding individual CNNs' results is adopted. In this paper, we have adopted the result aggregation method by voting.

In terms of the number of CNNs used by ensemble classifiers, for the two classes of identification problems, the number of CNNs is usually an odd number, so that when the vote is not given, there are two kinds of votes, which cannot determine the problem. However, it does not apply to multiclass identification problem because the class identification problem uses even odd numbers of CNNs, it is sometimes impossible to get the majority. For example, an ensemble classifier of seven CNNs is used to identify a three class classification problem. In terms of as unknown input image, assume that the result of three CNNs belongs to the first class, three belongs to the second class and one belongs to the third class. At this point, the first class is the same as the second class, and the final result cannot be determined. To solve this problem, our solution is to determine the final result according to the probability value of the classification output of the CNNs.

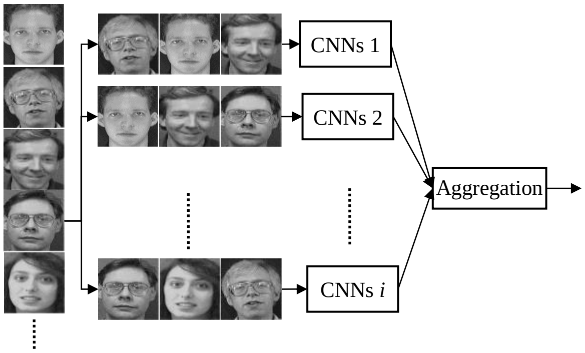

Figure 2: The Structure of Multiple-CNNs-based ensemble Classification of the Method 1.

3.2 Method 2

It is found through observing the face image database that, since the head pose is different and the distance from the lens is different, the image of the same person is slightly different in scale, which is easy to cause the error of face recognition. Therefore, the second method can effectively solve the problem of face recognition at different scales. The structure of the ensemble classifiers of the second method is shown in Figure 3. The structure of each CNNs is same as Figure 1. In the training phase, we generate different training sets in order to enlarge and reduce the size of each face image in the original face training set. Each training set has the same number of original training sets. For example, use five ensemble classifiers, namely enlarging by f and f^2 times and reducing by $1/f$ and $1/f^2$ times of the each face image size in the original training set to generate four new training sets. The new and original training sets are used to generate five CNNs respectively. In the test phase, in terms of unknown face images, the input images are enlarged and reduced using the same method. And then, input the image to the CNNs corresponding to each scale, calculate the individual results, and finally merge the results by voting methods.

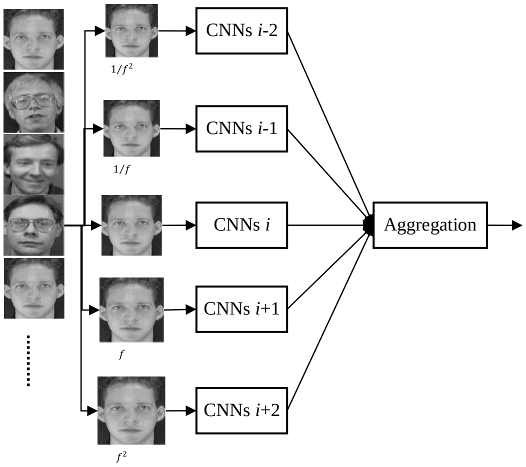

Figure 3: The Structure of the Multiple-CNNs-based ensemble Classification of Method 2.

4 Experiment Results

The following chapters will describe the experimental results and discussions of the above two methods. In this paper, we use the ORL face database [15] for face recognition experiments. This database of faces contains a set of face images taken between April 1992 and April 1994 at the Speech, Vision and Robotics Group of the Cambridge University. There are ten different images of each of 40 distinct subjects (as shown Figure 4). Therefore, the database includes 400 face images. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position. The size of each image is 92×112 pixels, with 256 grey levels per pixel. We will make each face image normalized to 28×28×1 pixels. 28, 28 and 1 respectively represent the face image width, height and channel, and the 400 face images are divided into 360 training sets and 40 test sets.

CNNs is composed by five layers. The first layer is convolution layer, which uses twenty filters with 5×5×1 pixels to generate feature maps with 24×24×20 pixels by convolution operation of the face images with 28×28×1 pixels. The second layer is pooling layer, which generates output feature maps with 12×12×20 pixels by non-overlapping maximum sampling of the input feature maps in the area with 2×2 pixels. The third layer is the second convolution layer, which uses 50 filters with 5×5×20 pixels to generate output feature maps with 8×8×50 pixels by convolution operation of the input feature maps with 12×12×20 pixels. The fourth layer is the second pooling layer, which generates output feature maps with 4×4×50 pixels from the feature maps with 8×8×50 pixels. The fifth layer is a fully connected two layer perceptron (MLP). The perceptron has 500 hidden layer nodes and 40 output layer nodes, representing 40 people in the face database. The output activity function uses the *softmax*( ) function, and the other activity functions use the *tanh*( ) function. The number of training drops is 200 times, the number of training batches is 40, and the learning rate is 0.01.

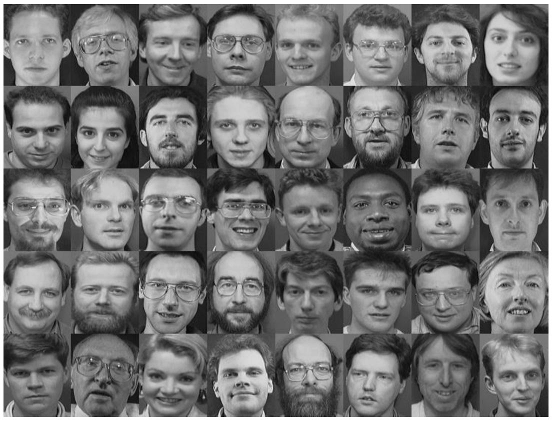

Figure 4: All Subjects of dataset.

Table 1 is the experimental result of ensemble classifier of the first method. The ensemble classifier uses a total of five CNNs and does the same experiment for five times. The results are shown from the second to the sixth list in Table 1. The individual test accuracy rate, the average value of the individual test accuracy rate and the test accuracy rate of ensemble classifiers of CNNs are recorded respectively in each list. The average value of the individual test accuracy rate of the five experiments are 88.0%, 83.5%, 91.5%, 85.0% and 82.0% respectively. The accuracy rate of the five experiments of the ensemble classifier are 90%, 85%, 95%, 90% and 90%, respectively, with an average of 90%. From table 1 we can see that the ensemble classifier does have a better accuracy, followed by the complementary role of individual CNNs. Take the third experiment as an example, 40 test images has 3, 4, 4, 5 and 1 images that has been identified incorrectly respectively from CNNs 1 to CNNs 5. The individual accuracy rate of CNNs are 92.5% (37/40), 90% (36/40), 90% (36/40), 87.5% (35/40) and 97.5% (39/40) respectively. The accuracy rate of the ensemble classifier is 95.0%, higher than the average value of 91.5% of individual accuracy rate, which means that different face images tested by individual CNNs for error identification makes the accuracy rate of ensemble classifier higher than the average accuracy rate value of individual CNNs. Therefore, it is proved that the way of ensemble does generate the effect of complementation.

We further remove the hair from each face in the database, as well as some of the background (shown in Figure 5), and then proceed to the same experiment as table 1, and the results are shown in table two. The average accuracy rate of individual CNNs in the 5 experiments are 73.0%, 64.5%, 70.0%, 80.5% and 77.0%, respectively. The accuracy of the ensemble classifier are 82.5%, 72.5%, 77.5%, 90% and 77.5%, respectively, with an average of 80%. From the results of table two, the ensemble classifier still achieves better test accuracy than individual CNNs. Secondly, the result of Table 1 is better than that of Table 2, which shows that background information and other facial information, such as hair style, beard, posture and facial contour have great influence on the accuracy of face recognition

Table 1 Five CNNs used by ORL Human Face database and the Test Accuracy Rate of the Method 1.

Trial	CNNs 1 (%)	CNNs 2 (%)	CNNs 3 (%)	CNNs 4 (%)	CNNs 5 (%)	Avg. (%)	Ensemble CNNs (%)
1	87.5	87.5	87.5	87.5	90.0	88.0	90.0
2	80.0	85.0	82.5	85.0	85.0	83.5	85.0
3	92.5	90.0	90.0	87.5	97.5	91.5	95.0
4	87.5	82.5	87.5	90.0	77.5	85.0	90.0
5	82.5	75.0	90.0	82.5	80.0	82.0	90.0
				Avg.			90.0

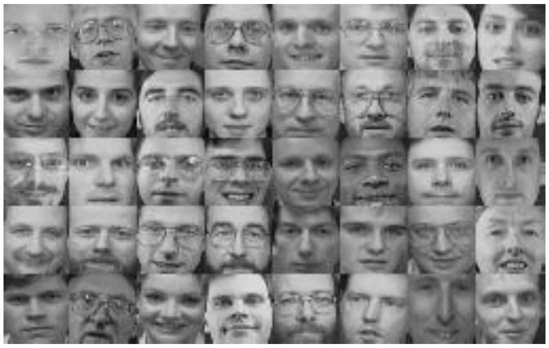

Figure 5: All Subjects of dataset with removing the part of Hair and Background.

Table 2: Five CNNs used by ORL Human Face Database that has been removed part of the Background and the Accuracy Rate tested by first Method.

Trial	CNNs 1 (%)	CNNs 2 (%)	CNNs 3 (%)	CNNs 4 (%)	CNNs 5 (%)	Avg. (%)	Ensemble CNNs (%)
1	77.5	70.0	60.0	72.5	85.0	73.0	82.5
2	72.5	70.0	67.5	55.0	57.5	64.5	72.5
3	52.5	72.5	72.5	75.0	77.5	70.0	77.5
4	82.5	75.0	77.5	82.5	85.0	80.5	90.0
5	70.0	77.5	77.5	80.0	80.0	77.0	77.5
				Avg.			80.0

Table 3: Five CNNs used by ORL Human Face database and the Test Accuracy Rate of the second Method. (Scaling factor *f* =1.1)

Trial	CNNs 1 (%)	CNNs 2 (%)	CNNs 3 (%)	CNNs 4 (%)	CNNs 5 (%)	Avg. (%)	Ensemble CNNs (%)
1	95.0	97.5	97.5	82.5	70.0	88.5	97.5
2	97.5	97.5	100.0	87.5	82.5	93.0	100.0
3	97.5	97.5	100.0	80.0	77.5	90.5	100.0
4	95.0	97.5	97.5	87.5	82.5	92.0	100.0
5	80.0	97.5	100.0	97.5	75.0	90.0	100.0
				Avg.			99.5

Table 3 is the experimental result of ensemble classifier of the second method. The results are shown from the second to the sixth list in Table 3. The average accuracy rate of individual CNNs in the 5 experiments are 88.5%, 93.0%, 90.5%, 92.0% and 90.0% respectively. The accuracy rate of ensemble classifier are 97.5% for one time 100.0%, for four times, with an average of 99.5%. From Table 3, we can get the following points. The first point is that ensemble has better accuracy rate. The second point is that the second method has a better accuracy rate than the first method. The third point is that the individual CNNs of the ensemble classifier of the second method also has complementary effects, and the erroneous identification of the face images can be used to obtain the correct identification results at different scales. Take the fourth experiment as an example, 40 test images has 2, 1, 1, 5 and 7 images that has been identified incorrectly respectively from CNNs 1 to CNNs5. The individual accuracy rate of CNNs are 95.0% (38/40), 97.5% (39/40), 97.5% (39/40), 87.5% (35/40) and 82.5% (33/40) respectively. The accuracy rate of the ensemble classifier is 100%, higher than the average value of 92.0% of individual accuracy rate, which has further explained this situation

Table 4: The Classification Result of Incorrect Classification Face Images in Testing Sets by using Ensemble Classifier of the second Method. (Scaling Factor *f* =1.1)

Correct classification	CNNs 1	CNNs 2	CNNs 3	CNNs 4	CNNs 5	Ensemble CNNs
2	10	2	2	2	2	2
4	4	4	4	4	18	4
5	20	5	5	5	5	5
6	6	6	6	6	34	6
8	0	0	8	8	8	8
10	4	10	10	10	10	10
11	0	11	11	11	11	11
23	23	23	23	23	26	23
24	24	24	24	13	13	24
29	7	7	29	29	29	29
30	30	30	30	30	5	30
33	7	33	33	33	13	33
37	37	37	37	37	34	37
38	38	38	38	38	21	38

We have listed the error classification result of 40 face images from CNNs 1 to CNNs 5 in the training set in another test. The first column of the table 4 is the correct classification of tested face images. The second to the sixth shows the classification result of CNNs 1 to CNNs 5 and the last column is the classification result of ensemble classifier with the incorrect classification marked with a background. Therefore, the number of incorrect classification are 7, 2, 0, 1 and 8 respectively from CNNs 1 to CNNs 5. From the table 4 we can see that the CNNs 3 with the original size has the correct classification, and the larger or smaller size of the training set CNNs, the more the incorrect classification. In addition, the face images with original size is tested appropriately and can be classified correctly in in slightly enlarged and reduced CNNs 2 and CNNs 4. For example, the 33 tested images which has been classified correctly in the table 4. However, in terms of the tested face images with original size, the smaller the size of CNNs, the more the correct classifications. The reason is that the face image whose original size is enlarged because of the reduced size can be classified correctly because of being complemented. For example, test face images that is correctly classified into 4, 6, 23, 24, 30, 37, and 38 in the table 4. On the contrary, in terms of the tested face images with relatively small original size, the larger size of the CNNs, the more the correct classifications. For example, test face images that is correctly classified into 2, 5, 8, 10, 11, and 29 in the table 4. Therefore, CNNs with different sizes can generate complementary effects.

5 Conclusion

In this paper, we have proposed two kinds of ensemble classifiers based on multiple CNNs. The difference between the two methods is that the training sets are generated differently. The first is to generate various different training sets by random sampling and then replacing of the original data. The second is to generate various different training sets by adjusting the image size of the original data and doing face recognition experiments using ORL face database. The average accuracy rate of the two kinds of ensemble classifiers are 90.0% and 99.5% respectively. The followings are the conclusions got from the experiment results. The first conclusion is that both two kinds of CNNs ensemble classifiers have better accuracy rate than the single CNNs classifier. Because individual CNNs of the ensemble classifiers can complement each other, the result got from merging makes the CNNs ensemble classifier better than single CNNs classifier. The second conclusion is that the accuracy rate of the second kind of CNNs ensemble classifier is better than that of the first kind. Because the second method takes the image size into consideration and the first kind of CNNs ensemble classifier generates training set by random sampling, some face images haven’t been comprehensively studied, resulting in relatively bad results. The third result is that in addition to facial features, other features such as contours, backgrounds, or other non-facial features also affect face recognition.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Science and Technology, Taiwan, R.O.C., under contract number MOST 105-2221-E-324-022.

REFERENCES

[1]. Y. Bengio, 2009, Learning Deep Architectures for AI, Foundations and Trends in Machine Learning, 1-127.

[2]. A Deep Learning Tutorial: from Perceptrons to Deep Networks, 2017, available online: http://www.toptal.com/machine-learning/an-introduction-to-deep-learning-from-perceptrons-to-deep-networks. (Accessed on September 26, 2017).

[3]. Brief Descriptions about Deep Learning, 2017, available online: http://onexinjuexing.blogspot.tw/2014/05/study-brief-descriptions-about-deep.html. (Accessed on September 26, 2017).

[4]. Deep Learning Net, 2017, available online: http://deeplearning.net. (Accessed on September 26, 2017).

[5]. Unsupervised Feature Learning and Deep Learning (UFLDL), 2017, available on: http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial. (Accessed on September 26, 2017).

[6]. Stochastic Pooling for Regularization of Deep Convolutional Neural Networks, Tech-talks TV, 2017, available on: http://techtalks.tv/talks/stochastic-pooling-for-regularization-of-deep-convolutional-neural-networks/58106. (Accessed on September 26, 2017).

[7]. Convolutional Neural Network, 2015, available on: https://en.wikipedia.org/wiki/Convolutional_neural_network. (accessed on 24 Dec. 2015).

[8]. MNIST data, 2015, Available on: https://en.wikipedia.org/wiki/MNIST_database. (Accessed on 24 Dec. 2015).

[9]. D. C. Ciresan, U. Meier, L. M. Gambardella and J. Schmidhuber, 2011, Convolutional Neural Network Committees for Handwritten Character Classification, In *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1250-1254.

[10]. D. C. Ciresan, U. Meier and J. Schmidhuber, 2012, Multi-Column Deep Neural Networks for Image Classification, In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3642-3649.

[11]. D. C. Cireşan and U. Meier, 2015, Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification, In *Proceedings of International Joint Conference on Neural Networks*, pp. 1-6.

[12]. V. Romanuke, 2016, Parallel Computing Center (Khmelnitskiy, Ukraine) Represents an Ensemble of 5 Convolutional Neural Networks Which Performs on MNIST at 0.21 Percent Error Rate, Retrieved 24 November 2016.

[13]. Y. Ren, L. Zhang and P. N. Suganthan, 2016, Ensemble Classification and Regression-Recent Developments, In *Proceedings of Applications and Future Directions*, pp. 44-53.

[14]. L. Breiman, 1996, Bagging Predictors, Machine Learning, pp. 123-140.

[15]. AT&T Laboratories Cambridge, 2017, available online: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. (Accessed on September 26, 2017).