BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 70(3), 2022, Article number: e141004
DOI: 10.24425/bpasts.2022.141004

Ensemble of classifiers based on CNN for increasing generalization ability in face image recognition

Robert SZMURŁO [1]*, **and Stanisław OSOWSKI** [2]

[1] Faculty of Electrical Engineering, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland
[2] Faculty of Electronic Engineering, Military University of Technology, gen. S. Kaliskiego 2, 00-908 Warszawa, Poland

Abstract. The paper considers the problem of increasing the generalization ability of classification systems by creating an ensemble of classifiers based on the CNN architecture. Different structures of the ensemble will be considered and compared. Deep learning fulfills an important role in the developed system. The numerical descriptors created in the last locally connected convolution layer of CNN flattened to the form of a vector, are subjected to a few different selection mechanisms. Each of them chooses the independent set of features, selected according to the applied assessment techniques. Their results are combined with three classifiers: softmax, support vector machine, and random forest of the decision tree. All of them do simultaneously the same classification task. Their results are integrated into the final verdict of the ensemble. Different forms of arrangement of the ensemble are considered and tested on the recognition of facial images. Two different databases are used in experiments. One was composed of 68 classes of greyscale images and the second of 276 classes of color images. The results of experiments have shown high improvement of class recognition resulting from the application of the properly designed ensemble.

Key words: CNN; ensemble of classifiers; face recognition; feature selection.

1. INTRODUCTION

The generalization of artificial intelligence systems represents the ability to adapt to the new, previously unseen testing data that come from the same distribution as learning data. The knowledge acquired in the learning process is automatically transferred to the new situation. The previous experience, which is similar in one or more ways to the new one is applied to develop decisions regarding the tested data samples. The learning process of the intelligent system (neural network or other intelligent solutions) is aimed at understanding the mechanism, based on which the learning data have been created [1–4].

In the typical situation, the network may not be sufficiently complex to learn this mechanism properly, or the population of learning data is too scarce and does not present sufficiently well the modelled process. This problem is especially difficult in deep learning structures, like convolutional neural networks, which contain millions of adapted parameters. According to Vapnik theory [4], the generalization ability strongly depends on the relation between the size of learning data and the complexity of network architecture. The higher this ratio, the better the expected generalization ability.

Many different techniques have been developed to improve the generalization. One of them is based on expanding the size of a training dataset by augmentation of data, like flips, translations, rotations, scaling, cropping, adding the noise, non-negative matrix factorization, creating synthetic images using self-similarity, application of GAN technique or variational autoencoder, etc. [3, 5]. However, in deep structures, where the number of parameters is counted in millions such a technique is of limited efficiency.

Another way to increase the generalization is the regularization of the architecture, achieved by such techniques, as modifying the structure by cutting some weight connections, weight decay, dropout, batch learning normalization, etc. [2].

An important method for increasing generalization capability is arranging many parallel solutions integrated into the ensemble [6, 7]. Different, independent team members, who look at the modelled process from a different point of view, form a so-called expert system, which makes it possible to generate a more precise decision for the data samples not taking part in the learning process.

However, to create a well-working ensemble we should provide the independence of its members [6, 7]. To achieve this goal different techniques are applied, such as random choice of learning data used in training of members of an ensemble, application of mini batches created randomly in the adaptation process of parameters, diversification of dropout ratio of learning data, differentiation of the type of units forming the ensemble, etc.

In this paper, we will analyse different strategies for forming an ensemble. The proposed ensemble is created based on many runs of the convolutional neural networks (CNN) [8, 9]. Among many existing deep neural networks, like Resnet, Inception, EfficientNet, etc., we have chosen Alexnet as the basic CNN architecture, since its application in this task was very efficient (relatively better accuracy and very efficient Matlab implementation). In our solution, we applied transfer learning

*e-mail: robert.szmurlo@pw.edu.pl

Manuscript submitted 2021-10-04, revised 2021-12-21, initially accepted for publication 2022-01-19, published in June 2022.

based on the ALEXNET architecture of CNN [10]. The full set of numerical descriptors of the size 4096 is taken from the fc6 layer of the network. These descriptors are subjected to selection for creating the diagnostic features. Five different selection methods: stepwise fit, nearest neighbour analysis, relief, Chi2 test, and minimum redundancy and maximum relevance methods [11-13] are applied. Since each method uses a different selection mechanism, the selected features form different sets. Each set represents the input signals to the classifier.

Two, carefully selected classical classifiers have been applied in forming the ensemble. One is the support vector machine (SVM) [14] and the second is the random forest of decision trees (RF) [15]. Both have the reputation of the best classical solution in a classification task. Associating them with 5 selection procedures allows the creation of 8 classifying members of the ensemble. Additionally, the softmax classifier built into ALEXNET [10, 16] is the next potential member of an ensemble. It is supplied by the randomly selected activation signals from fc7 layer according to the assumed value of the dropout ratio. In this way, in each run of image processing, the ensemble may be formed of up to 9 members. Moreover, the classification task may be repeated many times generating the results, which are integrated into the final verdict of an ensemble. Integrating them by majority voting into one final verdict of the whole ensemble can potentially increase the accuracy of the system.

Different arrangements of classification units have been created and tested. One is the composition of only one type of classifier (softmax, SVM, or RF), each combined with the specific set of diagnostic features. The second represents the combination of many types of classifiers in the ensemble. The results of these combinations will be compared, and the best choice used in the final experiments.

The numerical simulations will be performed on the task of automatic recognition of facial images using Matlab [16]. Automated facial recognition plays nowadays a very important role in many branches of our everyday life (biometric authentication of the person as a part of the inspection program that compares a face to their photo stored on the passport, identification of the identity of wanted individuals, etc.).

Face recognition is the topic that received the most attention from the research community. The studies proposed building efficient classification models concentrating on a different aspect of data processing, such as data augmentation, loss function design, or model design to combat adversarial attacks [17].

Many different solutions to the problem have been proposed in the past. In general, traditional methods attempted to recognize human faces by one- or two-layer representations, such as filtering responses, histogram of the feature codes, or distribution of the locally based descriptors [18]. The important approach used the linear principal component analysis (PCA) or linear discriminant analysis (LDA) and their different modifications (so-called Eigenface method) applied either directly or indirectly in the definition of diagnostic features, which are used as input signals to the classifiers [19, 20]. Good results of recognition have been obtained at the application of neural type classifiers (multilayer perceptron, support vector machine, etc.) or a random forest of decision trees [20].

Nowadays the best results of image recognition are obtained using deep learning (autoencoder, convolutional neural networks, etc.) [21]. Convolutional neural networks use a cascade of multiple layers of processing units for feature extraction and transformation. Different levels of abstraction, representing multiple views on the image are learned in this way. The internally developed numerical descriptors of the image are well correlated with the recognized classes, and thanks to this great improvement of the class recognition accuracy are possible.

The most important advantage of the deep approach is a combination of two basic steps (generation of numerical descriptors and final classification) in one common architecture. Thanks to this the process is very simplified since the most important and difficult stages of image processing are done automatically by the system.

The efficiency of the proposed solution is studied by recognizing facial images representing different people (treated as classes). Two databases are used for the numerical experiments. The first one consists of 68 classes of images represented in greyscale, and the second one consists of 276 classes of colour images [22]. The results are presented when applying each classifier, including the average of their actions, and when integrating their results into a differently organized ensemble. The experiments have confirmed a very high improvement in class recognition resulting from the application of many classifiers integrated into the ensemble.

2. ARCHITECTURE OF CLASSIFICATION SYSTEM

The proposed solution is based on the application of CNN as a workhorse. The set of locally connected convolutional layers generates the numerical descriptors of the input image. These numerical descriptors, subjected to various preprocessing methods, form the diagnostic features for the classification units that constitute the ensemble.

2.1. convolution neural network

The convolutional neural network used in the system is built based on ALEXNET architecture [10, 16] presented below

1 'data' Image Input 227x227x3 images with 'zerocenter' normalization
2 'conv1' Convolution 96 11x11x3 convolutions with stride [4 4] and padding [0 0]
3 'relu1' ReLU
4 'norm1' Cross Channel Normalization with 5 channels per element
5 'pool1' Max Pooling 3×3 max pooling with stride [2 2] and padding [0 0]
6 'conv2' Convolution 256 $5 \times 5 \times 48$ convolutions with stride [1 1] and padding [2 2]
7 'relu2' ReLU
8 'norm2' Cross Channel Normalization with 5 channels per element
9 'pool2' Max Pooling 3×3 max pooling with stride [2 2] and padding [0 0]
10 'conv3' Convolution 384 $3 \times 3 \times 256$ convolutions with stride [1 1] and padding [1 1]
11 'relu3' ReLU
12 'conv4' Convolution 384 $3 \times 3 \times 192$ convolutions with stride [1 1] and padding [1 1]
13 'relu4' ReLU
14 'conv5' Convolution 256 $3 \times 3 \times 192$ convolutions with stride [1 1] and padding [1 1]
15 'relu5' ReLU
16 'pool5' Max Pooling 3×3 max pooling with stride [2 2] and padding [0 0]
17 'fc6' Fully Connected 4096 signals (numerical descriptors)
18 'relu6' ReLU
19 'drop6' Dropout 50% dropout
20 'fc7' Fully Connected K neurons % K set individually by the user
21 'relu7' ReLU
22 'drop7' Dropout 50% dropout
23 'fc8' Fully Connected M neurons
24 'prob' Softmax
25 'output' Classification results of M classes

It contains five locally connected convolution layers and three fully connected layers. In the standard application of CNN, the images of the last locally connected layer are flattened to the vector form and represent 4096 numerical descriptors used as the input signals to the softmax classifier. It is composed of K hidden neurons of rectified linear unit (ReLU) activation and an output layer of M neurons representing the recognized classes. In our solution, this network will represent one member of the ensemble.

2.2. Other potential classifiers in ensemble

The other members of the ensemble will be created based on two different classification units: the support vector machine of Gaussian kernel and random forest of decision trees. Both classifiers will be supplied by the specially selected numerical descriptors formed from signals of the fc6 layer of the CNN.

The support vector machine [4,14] is regarded as the most efficient classical classifier. It is a supervised classification model strictly associated with a very special learning algorithm developed by V. Vapnik. The learning procedure constructs a hyperplane in a high-dimensional space providing a good separation between classes of data, characterized by the largest distance between the nearest training data of the classes (so-called margin of separation). Nonlinear mapping of the set of original vectors $\mathbf{x}$ into the hyperplane using a kernel function $K(\mathbf{x}, \mathbf{x}_i)$ allows much better discrimination between the data of two opposite classes that are not convex in the original space.

Thanks to this the SVM classifiers perform very well (good generalization ability) in difficult high-dimensional classification problems at a relatively small population of learning data. The SVM of the Gaussian kernel was used in our solution. The regularization constant C and Gaussian kernel width have been adjusted in an introductory stage by repeating the learning experiments for the set of their predefined values and choosing the best one based on the validation data set.

The Breiman random forest [15] represents a special ensemble of decision trees, that operates by constructing many decision trees at training time and outputting the class pointed by their majority. The very good generalization ability of the classifier is obtained by applying randomness in selecting the learning data, as well as using the limited set of randomly selected features chosen in each node of the tree.

3. FEATURE SELECTION PROCEDURES

Feature selection methods play an important role in the presented solution. They work with the set of 4096 numerical descriptors generated in the fc6 layer of the CNN. The crucial point in this phase is to provide the independent operation of the selection technique. This is achieved here by selecting methods that rely on different data validation mechanisms. There are many selection methods based on different principles, like statistical hypotheses, correlation principle, minimum redundance-maximum relevance, the distance of neighbouring samples, genetic algorithms, decision trees, etc. As a result of their application different sets of chosen descriptors can be selected. Our preliminary experiments have shown, that increasing their number beyond some limit does not sufficiently increase the efficiency of the system, however, results in slower operation. Based on this observation we have limited their number to five, carefully selected methods. The following methods have been used: stepwise fit, the nearest neighbour analysis, the relief test, the Chi2 test, and the minimum redundancy-maximum relevance criteria [11–13, 23].

Stepwise fit (SWF) selection [16] is based on adding and removing variables from the set of features. The continuous process of adding or removing variables to the selected set of features is controlled by checking if some variables from the set can be deleted without significantly increasing the error of classification. The selection procedure terminates when the actual quality measure of the classification results is maximized, or when the available improvement from step to step falls below some critical value.

The nearest neighbour analysis (NCA) is a special method of selection [13] based on the application of the K nearest neighbours (KNN) classifier. In the process of selecting the winners, the distances between input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are subjected to scaling and in the case of N-dimensional feature vectors this distance is defined in the form

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{N} w_l^2 \left| x_{il} - x_{jl} \right| . \tag{1}$$

The parameter w_l, which is assigned to the l-th feature, is subject to adjustment in the learning procedure. The rank of the variable in the set depends on this value.

The relieff (Relf) is another method of selection used in our solution [12]. In this method, each data instance represented by the actual feature vector $\mathbf{x}$ looks for the closest instances from each class in the database. The "nearHit" represents the closest same-class and the "nearMiss" – the closest different class. The i-th component of the vector $\mathbf{x}$ is associated with the weight w_i, subject to adaptation [24].

$$w_i := w_i - (w_i - \text{nearHit}_i)^2 + (w_i - \text{nearMiss}_i)^2 . \tag{2}$$

The weight w_i decreases if it differs from that near feature of the same class more than nearby instances of the other class and increases in the reverse case. After n iterations, each weight is normalized and represents the relevance vector. The selected set is composed of features of values greater than the assumed threshold.

In the minimum redundancy and maximum relevance (MRMR) selection method [23], we look for the minimum set of features which are highly "correlated" with class and at the same time the least "correlated" within themselves. In this method the "correlation" is replaced by the statistical dependency between variables and the mutual information is used to quantify this dependency. In this sense, the MRMR tends to maximize the dependency between the joint distribution of the selected features and the classification variable.

The last method of selection will apply the chi2 test [16]. It is used to test the independence of two events. Chi2 test measures how expected count E and observed count O of two variables (here feature and target class) deviate from each other. If the observed count of the feature is independent of the target (class) the score of chi2 is close to the expected count and the chi2 value is very small. So, a high chi2 value indicates that the

hypothesis of independence is incorrect (the feature is strictly correlated with the class). In other words, the higher the chi2 value the class is more dependent on the feature, and it can be selected for the model.

As it is seen the applied selection methods have relied on different principles and hence their results are expected to be highly independent.

3.1. Ensemble of classifiers

Ensemble of classifiers is composed of a set of classification units, whose individual predictions are combined in some way to classify the new examples. Different fusing methods are applied in practice: majority voting, dynamic voting, Bayes rule, Kulback-Leibler principle, etc. Their complexity differs a lot, however, the fusion results are still comparable. Therefore, we have decided to apply the least complicated majority voting, which is most often exploited in the classification tasks.

The base classifiers differ in the algorithm used, hyperparameters, representation or the training set, etc. To provide proper operation of ensemble its members should be of comparable accuracy and act independently. The independent operation of classifiers forming the ensemble may be provided in many ways:

- The most typical is to apply different mechanisms of deciding on a classification. It is achieved by applying different types of classification principles, for example, neural network, support vector machine, Bayes classifier, softmax, decision tree, etc.
- Also important is the application of different sets of diagnostic features selected from the set of all available numerical descriptors defined in the analysed process. This can be achieved by applying various selection procedures, which rely their operation upon different principles.
- Good results of independence are achieved also by differing the contents of the learning samples used in teaching the classification members of the ensemble. Usually, it is done by selecting randomly the learning smaller subsets from the whole available data set. This method is widely explored in a random forest of decision trees.
- The common practice is also the application of different hyperparameter choices in the applied classifiers. This may refer to a different number of hidden layers and the number of units in each layer of the multilayer classifiers or a varied dropout ratio in the case of the softmax classifier.

In our solution, we will apply and investigate all these mechanisms and compare the results of their application. The members of the ensemble will be selected from three classifier types: SVM, RF, and softmax. Moreover, we will investigate the effect of reducing the number of selected numerical descriptors. These descriptors are generated in the fc6 layer of CNN and then are subjected to selection by applying various selection mechanisms. Five selection procedures presented in the previous section (Stepwise fit, NCA, relieff, MRMR, and chi2) act on 4096 numerical descriptors taken from fc6 of the ALEXNET. Thanks to different mechanisms of selection, different sets of features are generated. Combining their results with two classifiers results in 8 different potential members of the classification system. The system may be supplemented by the softmax classifier built directly into the ALEXNET [10,16]. In this way, the complete ensemble system will contain a maximum of 9 members. Different arrangements of these classifiers are tried and their efficiency in class recognition will be compared.

Figure 1 presents the general structure of the proposed ensemble classification system. It is the maximum possible architecture tried in the experiments. By omitting some units, we can

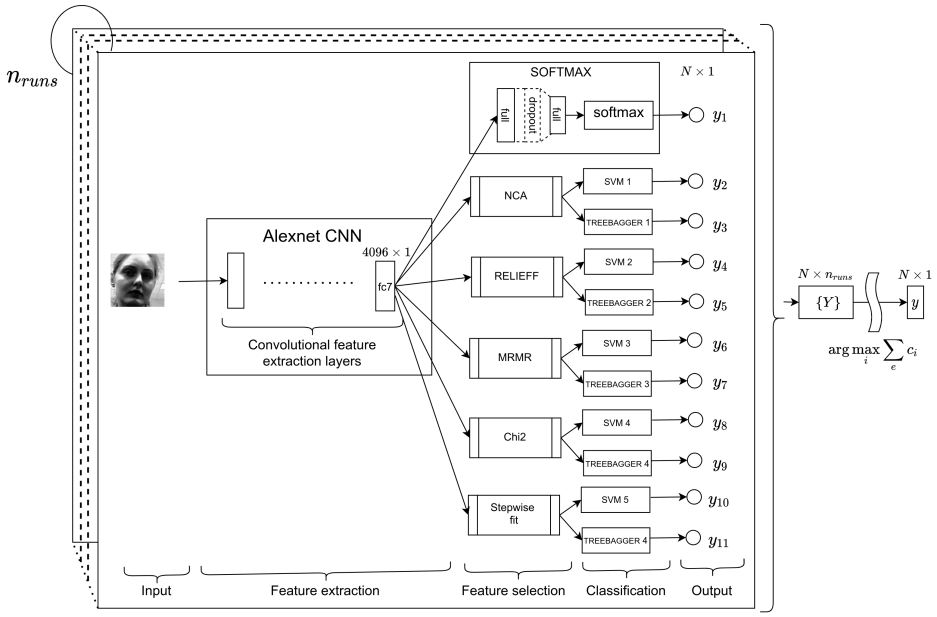

Fig. 1. CNN-based ensemble structure used in experiments. The pre-trained ALEXNET delivers the set of 4096 numerical descriptors of the data. They are subject to 5 selection procedures delivering a much smaller number of descriptors as the diagnostic features. These features represent the input signals to SVM and RF. The 11th classifier is built directly in the CNN structure and used the softmax method in classification. The ensemble structures can be formed by applying only some selected classifiers and omitting the others

investigate different arrangements of them and compare their efficiency in class recognition. All investigated architectures are run many times, delivering the results for fusion in developing the final verdict of class membership of the testing samples.

The condition of independent operation of ensemble members presented in Fig. 1 is achieved in various ways.

- The classifiers are trained based on the numerical descriptors taken from the fc6 layer of ALEXNET. Application of different selection procedures (5 of them are used in experiments) results in different reduced sets of diagnostic features. They form highly independent sets of input attributes applied to the classifiers.
- In the case of the softmax classifier, the structure of a fully connected network was varying, by applying a different number of hidden neurons in the fc7 layer (the population of neurons was changing from 700 to 900). Additionally, a different dropout ratio, changing from 0.3 to 0.6 was applied.
- Three different types of classifiers have been used: the softmax, SVM of Gaussian kernel, and random forest composed of a varied number of decision trees in each case.
- The learning stages of classifiers were performed using different, randomly selected sets of learning and validation images, keeping the testing data the same in all runs. The results of each run are combined in the ensemble results.

Since all systems have applied randomness in all phases of their learning operation (random choice of learning and validation data, random initialization of internal parameters, randomness included in some selection procedures, it was possible to increase the ensemble population by repeating the learning and testing processes few times, keeping the same set of testing data in each run. In this way, the integration of testing results was done on all results of individual runs of the classification systems.

4. DATABASES USED IN EXPERIMENTS

The developed classification system was tested using two different databases. The first database contained the faces of 68 persons [20,24], which are treated as 68 classes subject to recognition. Each class was composed of 20 photographs of the same person made in very different poses at varying lighting conditions. The size of the original images in all cases was the same and equal 100×100.

Figure 2 presents some chosen set of original images taking part in experiments. The same persons are photographed in different poses and at varying illumination. Some images show the face with glasses and some without glasses. The faces are shown in different scales, representing either full-face or only a limited part of it. Therefore, they represent a large variety of poses and viewpoints. The significant differences among the representatives of the same family of persons are visible.

The second database, called MUCT [22] is much larger (276 classes of images). Each class is represented by approximately 15 colour samples taken in a different position. The size of the images was the same and equal 480×640. The database consists of 3755 face images of different 276 subjects. Different

Fig. 2. The examples of face images represent 3 classes of data. Each column represents the images of persons belonging to the same class

poses were obtained thanks to the application of five different cameras at the same time. Additionally, everyone was photographed with 4 different lighting sets.

The database provides a diversity of lighting, age, and ethnicity of the people taking part in experiments. The images are taken in frontal and three-quarter views, different lighting sets, manual landmarks. The typical examples of images showing the variety of ethnicity of samples in the database are depicted in Fig. 3.

This time the images represent photos of the persons including not only faces but also the upper part of their body. The images seem to be easier in recognition, since the differences among class representatives concern many factors, like the colour of the skin, dress, hair, and ethnicity characteristics. Moreover, samples representing the same class are taken more similarly, showing only frontal and three-quarter views (see for example the last row of Fig. 3).

Fig. 3. The exemplary representatives of the MUCT database. The first two rows show persons of different ethnicity at varied lighting conditions. The last row depicts the exemplary differences in the presentation of the same person

5. NUMERICAL RESULTS OF EXPERIMENTS

The results of numerical experiments will be presented for two, presented above databases representing greyscale and colour images. Experiments aim to check how the inclusion of many independently working classifiers arranged in the form of an ensemble affects the quality of results. This quality will be assessed based on the accuracy, sensitivity, and precision in recognizing the class. The statistical results of different arrangements of an ensemble will be compared to the individual results of its members.

5.1. Results of numerical experiments for grey-scale images

The database was split into teaching parts (70% of randomly selected samples representing all classes) and the remaining 30% of the data representing the testing part (common to all classifiers in the ensemble). The learning part was once again randomly split into learning (80%) and validation part (20%). The cross-entropy formulation of the cost function was applied. ADAM learning algorithm with an initial learning rate of 1.3e-4 was applied in teaching. The mini-batch size was equal to 10 and the number of epochs was limited to only 20. The typical validation accuracy at the application of the softmax learning algorithm achieved in training on the validation set achieved a value around 98%.

The signals from the fc6 layer representing 4096 numerical descriptors of the images form the basis for further processing. These signals will be subjected to many different operations in the classification stage of the algorithm.

The quality measures presented here will show the results of different arrangements of the ensemble. The first investigated ensemble applied only a softmax classifier and was based on the results of 10 runs of experiments. The independence of the results of each run has been provided by different random selections of learning and validation data, various dropout ratios in softmax learning, and different number K of hidden units (random choice around mean of 800).

The important problem is to find the optimum number of components of the ensemble. Big data dimensions, a small population of databases, and limitations of available resources in terms of time and memory represent the most important issues and decisions of the best ensemble size. Therefore, the first experiments have been directed to estimate the proper choice of the ensemble.

Table 1 depicts such results, related to the quality measures corresponding to different statistics, including accuracy of the ensemble (ACC_{en}), mean of accuracy in individual results (ACC_{mean}), a median of individual results (ACC_{med}), maximum (ACC_{max}), and standard deviation of results (std). They correspond to the testing data not taking part in training and show the results at changing the number n of ensemble members.

Table 1

The statistical results in 68 class recognition at the application of ensemble applying different numbers of softmax classifiers (testing data only)

n	ACC_{en}	ACC_{mean}	ACC_{med}	ACC_{max}	std
5	96.3%	90.7%	90.9%	93.4%	1.84%
10	97.3%	91.4%	92.0%	92.9%	2.53%
55	97.8%	91.3%	91.9%	96.1%	2.39%

Based on these results and taking into account the calculation time, which is proportional to the number of members, it seems reasonable to limit the number of ensemble members to 10. Increasing this number to higher values has resulted in a non-significant increase in accuracy, compared to a sharp increase in calculation costs.

The advantage of the application of many members and integrating their results into one final verdict by majority voting is evident. The accuracy of the ensemble is much higher than the average of non-integrated individual members and this difference is very high (97.8% of ensemble versus 91.3% of the mean in the case of 55 members of ensemble). The interesting point is the increase of accuracy over the best individual member (97.8% of ensemble versus 96.1% of the best individual result).

The next experiments have been directed to apply different types of classifiers to form an ensemble. Two solutions have been tried: the SVM and random forest of decision trees. To achieve good generalization ability of these classical classifiers, we must limit the number of their input signals. Since the existing feature selection methods apply different mechanisms of feature assessment their results might differ. Therefore, few feature selection algorithms based on different principles have been used in experiments. They aimed to find the most important numerical descriptors among the 4096 generated by CNN and used them as input signals to the classifiers. Five different selection procedures have been applied (stepwise fit, NCN, relieff, MRMR, and chi2) and combined either with SVM or RF in an ensemble. The number of selected features was limited to only 2000 among the highest rank descriptors. The results of the application of individual classifier (SVM and RF) associated with five sets of preselected features (5 results in each run) for 10 runs of experiments are shown for the testing data in Table 2.

Table 2

The statistical results in 68 class recognition problems at the application of ensemble composed of either SVM or random forest classifiers

Classifier	ACC_{en}	ACC_{mean}	ACC_{med}	ACC_{max}	std
SVM	96.3%	90.9%	90.9%	94.4%	1.6%
RF	83.3%	76.3%	76.1%	80.6%	2.6%

The results show the advantage of SVM over the random forest. The RF is much more sensitive than SVM to the size of the population of images. However, both are inferior to the classic softmax classifier application. This is the result of a very small number of learning resources, which is a crucial point in obtaining good generalization ability of classical classifiers.

It is interesting to examine how increasing the number of units in an ensemble by combining Softmax with the other combinations of classifiers affects the performance of the classification system. The statistical results of such experiments for the same test data that did not participate in the training are shown in Table 3.

Table 3

The statistical results in 68 class recognition at the application of ensemble combined from different types of classifiers: softmax, SVM, and random forest at 10 runs of experiments

Classifier	ACC_{en}	ACC_{mean}	ACC_{med}	ACC_{max}	std
Softmax	97.3%	91.4%	92.0%	92.9%	2.5%
Softmax+ SVM	96.4%	91.3%	91.4%	94.8%	1.5%
Softmax+ RF	87.3%	81.0%	79.3%	95.1%	5.6%
SVM+RF	93.9%	85.0%	85.4%	94.6%	6.9%
Softmax+ SVM+RF	92.4%	84.0%	87.9%	93.6%	7.5%

Figure 4 presents the graphical comparison of the accuracy of different forms of ensemble arrangements in recognition of grayscale images. It is evident, that the best results correspond to the structure applying softmax classifiers.

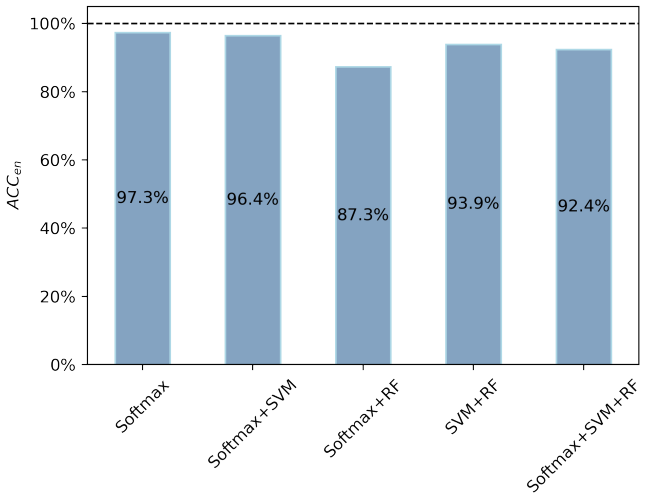

Fig. 4. The comparison of the accuracy of 68 class recognition achieved by different arrangements of the ensemble. The best results are due to the application of softmax as the basic unit of the ensemble

These results show the advantage of the softmax strategy over classical approaches to classification at a very small number of learning samples. The small size of the database results in a significant decrease of generalization ability in the case of classical classifiers, for which the proper number of learning samples is a crucial point. The deep learning algorithm was primarily tailored for such a case and thanks to this CNN is less sensitive to the limitation of learning resources.

The best choice of the ensemble was assessed also from the point of view of other quality measures. They included the recall, precision, and F1 coefficient. Recall R (often called sensitivity) of the class recognition is the fraction of the relevant instances that have been retrieved from all class samples in the testing stage of the classification system. On the other side, the precision P of the class recognition is the fraction of relevant instances among the retrieved instances. F1 measure is the ratio of the product of R and P related to their average value. For i-th class, this measure is defined as [11]

$$F1(i) = 2\frac{R(i)P(i)}{R(i)+P(i)}. \tag{3}$$

Figure 5 presents the values of $R(i)$, $P(i)$, and $F1(i)$ for 68 classes considered in experiments. These results correspond to the best ensemble built at the application of softmax-based units. As it is seen most classes have been recognized perfectly (100% of the quality measures). The representatives of only 10 classes among 68 subjected to recognition have been recognized with some errors: one misclassified representative in 9 classes and 2 misclassifications in one class (results of Recall in Fig. 5).

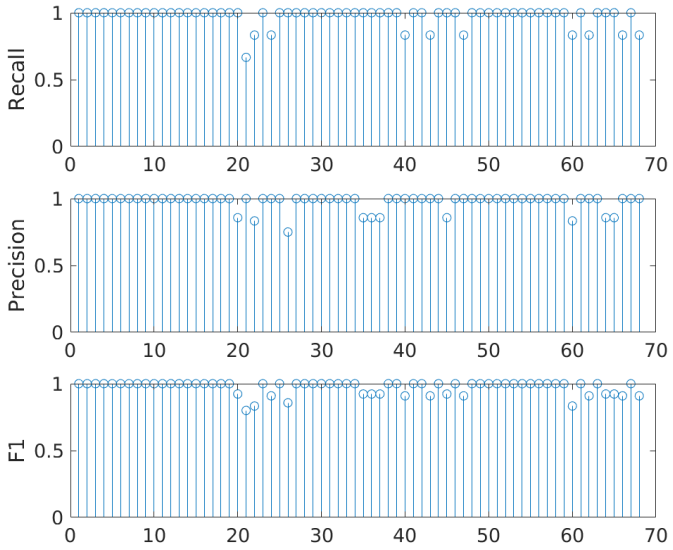

Fig. 5. The plot of recall, precision, and F1 measures for all 68 classes of greyscale images subject to recognition

5.2. Results of numerical experiments for color images

The next experiments are done using the MUCT database representing 276 classes of colour images [22]. Each person forming the class was represented by 15 sample images. The experiments were arranged in the same form as for greyscale images. 70% of samples representing each class were used for training and 30% left for testing purposes. The testing set was the same in each arrangement of an ensemble, while the learning and validation data were randomly selected from the training set in the runs.

The arrangements of the ensemble were organized similarly to the previous experiments with the greyscale images. They represent the structures applying only individual classifiers (softmax, SVM and RF), the combination of two classifiers in an ensemble (softmax+SVM, softmax+RF, SVM+RF), and the system applying all three classifiers (softmax+SVM+RF). The statistical results concerning the testing data are presented in Table 4.

Table 4

The statistical results in 276 class recognition problem at the application of ensemble combined from different types of classifiers: softmax, SVM, and random forest

Classifier	ACC_{en}	ACC_{mean}	ACC_{med}	ACC_{max}	std
Softmax	100%	98.9%	99.0%	99.7%	0.55
SVM	99.3%	97.9%	97.8%	98.8%	0.59
RF (no Stepwise – 40)	94.5%	85.4%	86.0%	89.9%	2.56
Softmax+ SVM	99.9%	98.9%	99.1%	99.8%	0.59
Softmax+RF	95.5%	86.0%	84.9%	99.8%	6.72
SVM+RF	99.5%	90.0%	92.5%	98.8%	8.45
Softmax+ SVM+RF	99.3%	91%	97.8%	99.5%	8.56

The best results were obtained for the ensemble based on the softmax classifier (the combination of results obtained in many runs with a different arrangement of the softmax structure (number of hidden units, different dropout ratio, etc.). The ensemble achieved 100% accuracy, although the individual runs resulted in different accuracies. Note that the best member of the ensemble achieved 99.7% accuracy, while the mean was 98.9%.

The interesting thing is the very high accuracy in detecting 276 classes, although the data population was rather small (15 representatives of each class). In our opinion, two factors are here of great importance. First, the individuals forming the classes were photographed in a similar position, so the diversity within the class was relatively low. Second, the MUCT database contains classes represented by very diverse individuals, varying by ethnicity, skin colour, etc. This fact increases the variance between classes, which is very helpful in classification.

We also studied the influence of colour on the recognition process. Some additional experiments were performed for images converted to grayscale. The results were only slightly worse (the overall accuracy of the best ensemble arrangement dropped from 100% to 99%).

6. SUMMARY AND DISCUSSION

The paper has studied different forms of creating the ensemble of classifiers to provide the best generalization ability of the classification system in the recognition of facial images. The proposed system was based on the deep neural network of the ALEXNET structure.

Experiments aimed to develop the optimal ensemble system, which provides the highest accuracy of face image recognition. The main role of the CNN structure is delivering the large set of numerical descriptors of the analysed images. They are generated in the last locally connected convolution layer of CNN and flattened to a form of vector.

These descriptors are subject to a few different selection mechanisms, responsible for creating the optimal subsets of class discriminating features. Thanks to the application of different selection methods, they provide independence of applied classifier operation, which is an important condition in creating the ensemble. However, it should be remembered that the selected sets are not globally optimal. Their optimality depends on the applied mechanism of selection, which may vary from method to method.

The developed sets of features are combined with three efficient classifiers: softmax, support vector machine, and random forest of the decision tree, which are employed in the final stage of image recognition. Their results are integrated into the definitive verdict of the ensemble. Different forms of arrangement of the ensemble have been considered and tested using two different datasets. One composed of 68 classes (persons) was in greyscale and the second representing 276 classes was in colour RGB representation.

The results of experiments have shown, that the most efficient in operation is the ensemble composed of softmax classifiers, employing different values of hyperparameters. Its effi-

ciency was significantly better than the mean of individual operations of the ensemble members. Moreover, the accuracy of the integrated ensemble was higher, than the result of the best individual. This is confirmed for both, greyscale and colour images. It is an interesting observation, that the performance of the proposed ensemble is practically not dependent on the number of the recognized classes. The generalization ability of the ensemble was very good despite a very small number of samples representing different classes.

It should be noted that such an ensemble arrangement based only on Softmax is the best for the actual database of face images. The other arrangements based on a mixture of Softmax, SVM, and RF were slightly worse, mainly due to the very small number of learning samples, which were insufficient for proper learning of the classical classifiers. The results may be different if the population of the database is larger.

The comparison of our results with other papers will be done on the example of the internationally recognized MUCT database available free on the internet.

The MUCT database was used in the past by many other researchers. For example, the paper [24] has declared the recognition accuracy changing from a minimum of 88.69% to an average of 98.3% (maximum). Their results were obtained for only 199 classes selected from the total of 276 classes in MUCT.

The paper [25] has investigated the influence of low resolution of images on the reliability of face detection and recognition and applied it to the whole MUCT database. The best recognition efficiency obtained for the original images of the resolution 256×256 was below 80%.

The paper [26] has investigated different arrangements of SVM networks for recognizing the face images in the MUCT database. The best-declared accuracy obtained for the set of 1512 images (1212 learning and 300 testing) selected from the database was equal to 93.7%.

Our best result obtained for all 3755 images is almost 100% in recognition for all 276 classes existing in the MUCT database.

The paper [27] has investigated the MUCT database in application to pose-invariant face recognition in robotics. Out of all original images of the database they have pre-selected only 1221 faces used in experiments. Only images of persons without glasses and with their mouths closed were used. The accuracy was reported in the form of a plot composed of a false acceptance ratio (FA) in the horizontal axis and a false rejection ratio (FR) in the vertical. At $FA = 1.1 \times 10^{-4}$ the value of $FR = 0.03$, while at $FA = 0.01$ the $FR = 0.05$.

The results of numerical experiments have shown, that although significant progress has been made in recent years with the deep learning approaches to face recognition there is still space for improving the efficiency of the class recognition systems, especially when a limited amount of data is available. The presented approach is of universal application and can be used in classification problems associated with other types of images.

REFERENCES

[1] Poggio and Q. Liao, "Theory I: Deep networks, the curse of dimensionality," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 66, no. 6, pp. 761–773, 2018, doi: 10.24425/bpas.2018.125924.

[2] Q. Zheng, M. Yang, J. Yang, Q. Zhang, and X. Zhang, "Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process," *IEEE Access*, vol. 6, pp. 15844–15869, 2018, doi: 10.1109/ACCESS.2018.2810849.

[3] P. Zhou and J. Feng, "Understanding generalization, optimization performance of deep CNNs," in *Proceedings of the 35th International Conference on Machine Learning (PMLR 80)*, Stockholm, Sweden, 2018.

[4] V. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[5] B. Swiderski, L. Gielata, P. Olszewski, S. Osowski, and M. Kołodziej, "Deep neural system for supporting tumor recognition of mammograms using modified GAN," *Expert Syst. Appl.,* vol. 164, pp. 1–10, 2021, doi: 10.1016/j.eswa.2020.113968.

[6] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley, New York, 2004.

[7] H. Bonab and F. Can, "Less is more: a comprehensive framework for the number of components of ensemble classifiers," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 14, pp. 2735–2745, 2018, doi: 10.1109/TNNLS.2018.2886341.

[8] I. Goodfellow, Y. Bengio, and A Courville, *Deep learning*, MIT Press, 2016.

[9] F. Chollet, *Deep Learning with Python*, Manning Publications Co., 2017.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems* – 1, 2012, 1097–1105, doi: 10.1145/3065386.

[11] P.N. Tan, M. Steinbach, and V Kumar, *Introduction to data mining*, Boston: Pearson Education Inc., 2006.

[12] R. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003, doi: 10.1023/A:1025667309714.

[13] W. Yang, K. Wang, and W. Zuo., "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, pp. 161–168, 2012, 10.4304/jcp.7.1.161-168.

[14] B. Schölkopf and A. Smola, *Learning with kernels*, Cambridge, MIT Press, MA, 2002.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 11, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] *Matlab user manual*, MathWorks, Natick, USA, 2021a.

[17] C.C. Loy *et al.*, "Editorial: Special issue on deep learning for face analysis," *Int. J. Comput. Vision*, vol. 127, pp. 533–536, 2019, doi: 10.1007/s11263-019-01179-z.

[18] M. Wang and W. Deng, "Deep face recognition: a survey," *Neurocomputing,* vol. 429, pp. 215–244, 2021, doi: 10.1016/j.neucom.2020.10.081.

[19] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 711–720, 1997, doi: 10.1109/34.598228.

[20] K. Siwek and S. Osowski, "Comparison of methods of feature generation for face recognition," *Przegląd Elektrotechniczny*, vol. 90, pp. 206–209, 2014, doi: 10.12915/pe.2014.04.49.

[21] M.M. Ghazi, H.K. Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2016, doi: 10.1109/CVPRW. 2016.20.

[22] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," *Pattern Recognition Association of South Africa* – database 2010.

[23] H.C. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI. 2005.159.

[24] B. Belavadi, K.V.M. Prashanth, G. Sanjay, and J. Shruthi, "Gabor Features for Single Sample Face Recognition on Multicolor Space Domain," *2017 International Conference on Recent Advances in Electronics and Communication Technology*, 2017, doi: 10.1109/ICRAECT.2017.23.

[25] T. Marciniak, A. Chmielewska, R. Weychan, M. Parzych, and A. Dabrowski, "Influence of low resolution of images on reliability of face detection and recognition," *Multimed. Tools Appl*, vol. 74, pp. 4329–4349, 2015, doi: 10.1007/s11042-013-1568-8.

[26] J.A.C. Moreano, N.B. La Serna Palomino, "Efficient Technique for Facial Image Recognition with Support Vector Machines in 2D Images with Cross-Validation in Matlab," *WSEAS Trans. Syst. Control,* vol. 15, pp. 175–183, 2020, doi: 10.37394/23203.2020. 15.18.

[27] M. Grupp, P. Kopp, P. Huber, and M. Ratsch, "A 3D face modelling approach for pose-invariant face recognition in a human robot environment," in *RoboCup 2016: Robot World Cup XX. RoboCup 2016. Lecture Notes in Computer Science*, S. Behnke *et al.* Eds., Springer, vol 9776, pp. 121–134, 2017, doi: 10.1007/978-3-319-68792-6_10.