PRZEGLĄD ELEKTROTECHNICZNY

Ukazuje się od 1919 roku **4'18**

Organ Stowarzyszenia Elektryków Polskich • Wydawnictwo SIGMA-NOT Sp. z o.o.

Invited paper

Krzysztof SIWEK[1], Stanisław OSOWSKI[1,2]

Warsaw University of Technology (1), Military University of Technology (2),

doi:10.15199/48.2018.04.01

Deep neural networks and classical approach to face recognition – comparative analysis

Abstract. *The paper presents application of the convolutional neural network (CNN) in face recognition. Data bases of faces have been represented by the visible and thermal infra-red images. The CNN is regarded nowadays as the most efficient tool in image analysis. This technique was applied to recognition of 50 classes of face images represented in visual and infrared imagery. This approach will be compared to the traditional approach relying on classical feature generation methods and application of support vector machine classifier. The numerical results of experiments performed on the face image data base will be presented and discussed.*

Streszczenie *Praca przedstawia porównanie metod rozpoznawania twarzy przy zastosowaniu konwolucyjnych sieci neuronowych (CNN) i klasycznego podejścia opartego na specjalistycznych metodach generacji cech diagnostycznych. Twarze są reprezentowane w postaci 2 rodzajów obrazów: widzialnego oraz w podczerwieni. Zbadano i porównano dwa podejścia do analizy obrazów. Jeden polega na zastosowaniu konwolucyjnej sieci neuronowej łączącej w jednym systemie generację nienadzorowaną cech diagnostycznych i klasyfikację. Drugie, klasyczne podejście, rozdzielające obie części przetwarzania. Generacja cech odbywa się poprzez zastosowanie specjalistycznych metod (tutaj PCA, KPCA i tSNE), a klasyfikacja wykorzystuje te cechy jako sygnały wejściowe dla oddzielnego klasyfikatora SVM. Wyniki eksperymentów numerycznych zostały przedstawione i porównane na bazie 50 różnych obrazów twarzy stworzonych w różnych warunkach oświetlenia i akwizycji.* ***Uczenie głębokie i podejście klasyczne do rozpoznawania obrazów twarzy - analiza porównawcza***

Słowa kluczowe*:* CNN, transfer learning, obrazy widzialne w podczerwieni, rozpoznawanie twarzy, transformacje danych, klasyfikacja.
Keywords: CNN, transfer learning**,** visible and infra-red imagery, face recognition, transformation of data, classification.

Introduction

The problem of face recognition is an important subject in image processing, since it has found large application in different solutions of safety systems. Two different forms of image acquisition have been most often used in practice: the visual (V) and infrared (IR) imagery. The visual cameras react on electromagnetic energy in the visible spectrum range from 0.4µm to 0.7µm, while sensors in the IR system respond to thermal radiation in the spectrum ranges from 0.7µm to 14µm. Moreover, the light in thermal IR cameras is emitted rather than reflected. The most important advantage of IR camera is its independence on illumination environment. The face detection, location and segmentation at varying lighting conditions are relatively easier than these in visual images [3,8,11]. However, there are also some disadvantages, such as loosing some details of the face, sensitivity to presence or absence of glasses, etc.

Irrespective of the acquisition method of the face images the most important point in recognition is the applied solution of the classification system. Traditional approach to this problem relies on characterization of the image by the set of numerical descriptors representing the input attributes to the classifier. These descriptors may be based on different principles. However, the most often used are the linear or nonlinear transformation techniques, like principal component analysis (PCA), kernel PCA (KPCA) and the stochastic neighbor embedding with a Student distribution (tSNE) [7], found as very useful tools in image preprocessing.

This paper will deal with application of deep learning strategy in image recognition. We have applied convolutional neural networks (CNN) regarded now as the most efficient tool in image processing [2,4]. CNN is a multilayer feedforward neural structure responsible for simultaneous generation of diagnostic features and classification. The first few locally connected convolution layers are responsible for the unsupervised generation of diagnostic features and the last fully connected layer represents the classifier, responsible for final recognition and classification.

The experiments have been performed on the data base composed of 50 classes, each represented by 20 face images of particular person. The images have been acquired using visible and infrared imagery. The acquisition has been done in different lighting conditions, different poses of persons and changing size of images. The results of CNN applications have been compared to the traditional approach using classical image preprocessing approach, applying PCA, KPCA and tSNE methods [5,7].

CNN approach to image recognition

Convolutional neural network is a multilayer feedforward structure, which performs at the same time two roles: the unsupervised generation of diagnostic features and classification. In contrast to traditional network of full connections between neurons in neighboring layers, it contains many hidden convolutional layers of local connections (the neurons in the next layer are connected to only small region in the previous layer). These layers perform the role of feature generation. Only the last one or two layers are fully connected and represent the classification unit. The typical structure of CNN applied to recognition of the classes of faces represented by the full original images is presented in Fig. 1. The data in convolutional layers are arranged in the form of 3-D tensor (horizontal and vertical dimensions of the image and the depths representing the succeeding images).

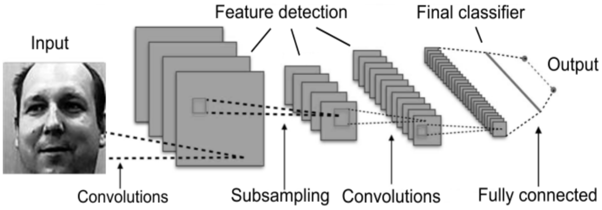

Fig. 1 The exemplary structure of CNN. The first 3 hidden layers composed of convolution and pooling sublayers represent symbolically the unsupervised feature extraction and the last part is a final classifier.

The convolutional layer realizes the linear convolution operation to the input represented by the pixels in the small reception field of the previous layer. This operation for image *I* and kernel function *K* is described by the following equation

(1) $Y(i,j) = I(i,j) * K(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$

The result of such operation is subject to the nonlinear processing described usually by the rectified linear unit (ReLU), generating output *y*(*x*) according to the formula

(2) $$y(x) = \begin{cases} x & for \quad x > 0 \\ 0 & for \quad x \leq 0 \end{cases}$$

In backpropagation learning of CNN this function is approximated in a smooth form

(3) $$y(x) = \ln(1 + e^x)$$

which has also the smooth derivative (required in backpropagation)

(4) $$\frac{dy(x)}{dx} = \frac{1}{1+e^{-x}}$$

The convolutional operation is performed on the pixels located in the reception field of the input image. This analysis is performed with the field moving along the image with the assumed stride. After normalization and pooling operation the new set of images forming the next layer is created. Regardless of the image size, the filtering process (weighted summing of the pixel intensity) is concentrated on the small masks representing pixels in the actual reception field of the analyzed image.

As a result only simple local processing of the images is needed. For instance at the mask size 5x5 the analyzing neuron has only 25 learnable parameters (weights), which are the same for every position of the moving mask. Thanks to this we avoid the problem of exploding gradient size in training the multilayer neural network by using backpropagation. The filtering process is traveling along the input image, creating the intensity values of the pixels, which form the resulting output image. The number of these images in each layer is equal to the analyzing neurons and defined by the user.

The CNN applies usually many hidden convolutional layers. Each layer is specializing in extracting the primitive features of the images (dots, crossing points, edges, etc.) in the succeeding layers, starting from the most abstract in the first one and ending in some complex combinations of them.

After processing image by the chosen number of locally connected neurons, we arrange the set of the reduced size images (tensors) of the last convolutional layer in the vector form, which represent the set of input attributes to the real classifier. They form the automatically extracted diagnostic features, which serve as the input attributes to the fully connected layers representing the final classifier of the system.

The most typical classifier in CNN is the so called softnet [2]. It is a simple one-layer classifier of the number of outputs equal to the number of recognized classes. Each output neuron is connected to all elements of the vector of input attributes. The weights are adapted in an usual way by solving the optimization process directed to minimization of the error function. The output signal $u_i(\mathbf{x})$ of each neuron is the weighted sum of input signals x_j (elements of input attribute vector **x**) and defined as

(5) $$u_i(\mathbf{x}) = \sum_j w_{ij}x_j + w_{i0}$$

The probability of membership of vector **x** to *i*th class is calculated using softmax function defined as [2]

(6) $$softmax(\mathbf{u})_i = \frac{\exp(u_i)}{\sum_{j=1}^{M} \exp(u_j)}$$

where *M* is the number of recognized classes. The largest value of softmax function dictates the class membership of the vector **x** corresponding to the actual image under classification. This form of classifier is very simple and at the same time found effective in the role of classifier.

The structure of CNN contains very large number of adjustable parameters. Therefore, learning process requires huge number of learning data and very long time. To counteract such situation the so called *transfer learning* is applied in practice, in which the user applies the initially pre-trained CNN structure. Such structures are trained at application of millions of images of arbitrary nature taken from internet. Actually, there are many such structures, like ALEXNET, ZFNet, GoogLeNet or VGGNet network, available in CAFFE repository [12]. The initially pre-trained network taken from this repository is subject to final training using the real data of the user. Thank to such approach the learning process is shortened to few minutes using GPU processor.

Classical approach to image recognition

In classical approaches to face recognition and classification, the image is first represented by the numerical descriptors, characterizing the structures of pixels in the most unique way for particular set of images belonging to the same class. In the case of images representing different classes the particular descriptor values should be as different as possible. Different preprocessing methods leading to various descriptor definitions are applied in practice. To the typical belong: principal component analysis, linear discriminant analysis, kernel PCA or stochastic neighbor embedding [1,7]. All of them reduce the size of the input image to the relatively small dimension of image descriptive vector.

As a result, the original image (the matrix converted row by row to the vector **x** of dimension *N*) is represented by the vector **y** of dimension *K*, much smaller than *N*. The PCA represents linear mapping **y=Wx** of the transformation matrix $\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K\right]^T$, defined on the basis of eigen-value decomposition (EVD) of the auto-covariance matrix [1], so called eigen-faces. In similar way the nonlinear kernel PCA is defined. The KPCA is just PCA performed on the nonlinear transformation of the vectors **x** [6,9].

In tSNE method we try to find the mapped elements of the reduced vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ representing the original high-dimensional data (vectors $\mathbf{x}_i$ and $\mathbf{x}_j$) of the image in a way to minimize a Kullback-Leibler divergence between the joint probability distribution p_{ij} in high-dimensional space and a joint probability distribution q_{ij} in the transformed (lower dimensional) space [10]. The value of p_{ij} represents probability that vector $\mathbf{x}_i$ is the closest neighbor to $\mathbf{x}_j$, while q_{ij} is the same measure for the transformed vectors $\mathbf{y}_i$ and $\mathbf{y}_j$. The transformation is aimed on finding the nonlinear mapping of the input vectors $\mathbf{x}_i$ which preserves the relative distances between the original vectors in a reduced space.

All these preprocessing methods lead to the representation of image by the limited number of diagnostic features. The features should represent the original vectors (face images) in a way providing the highest uniformity within the same class and highest differences for images representing different classes.

The numerical features created in this way form the input attributes to the classifier, responsible for the final recognition of classes. As the classifiers we have used here the Support Vector Machine of Gaussian kernel [6], which has the reputation of being the most efficient in classification problems. The hyperparameters of SVM (the regularization constant *C* and Gaussian kernel width) have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one on the validation data sets.

Data base of images

In our work we will compare the performance of the classification systems on the basis of two types of face representation: the visible and IR. The typical examples of these types of face images in different face arrangements (glasses and no glasses) and varying lighting conditions are presented in Fig. 2.

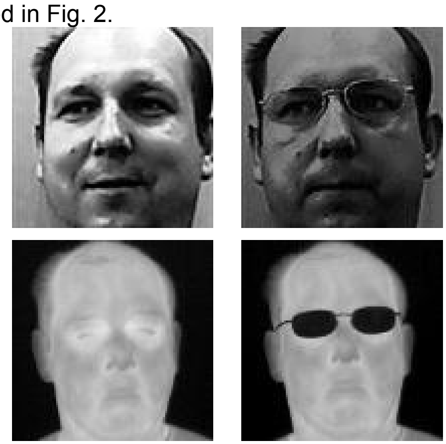

Fig. 2. Two pairs of visible and thermal IR images of the person with and without glasses and at different lighting.

Thermal IR imagery is nearly invariant to changes in ambient illumination, while the visible images are very sensitive to them. Therefore, in the case of IR imagery we expect reduction of the within-class variability as a result of different illumination in the acquisition process of face images. However, at the same time we note, that IR imagery results in loss of many significant details of the face, which might be significant in recognition of images belonging to different classes. This is especially true in the case of glasses, which cover the eyes.

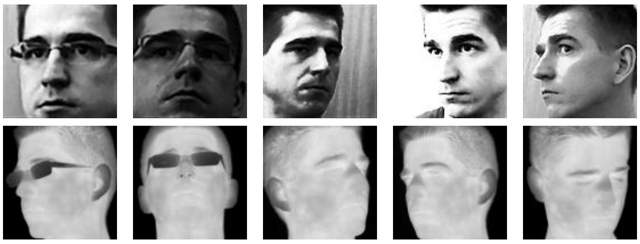

Fig. 3. The examples of diversity of face image acquisition of one person taking part in experiments: the upper row – the visual images, the lower row – infra red images.

The data base used in experiments was composed of the set of face images representing 50 classes of people (both men and women). Each class was represented by 20 individuals in different poses and illumination conditions. The same images have been acquired simultaneously in visible and infra-red forms. The size of original images in both cases was the same and equal 100×100.

The typical examples of images from the data base are illustrated in Fig. 3. They are presented in visual and infra-red imageries and differ by the size of the face, its position toward camera, presence or absence of glasses and also the background. Significant changes of intensity of pixel values are observed in the case of visible images. This is not the case in thermal IR representation. However, glasses occupying some part of the image cover important part of the face and may present difficulty in face recognition, especially in the IR representation.

Numerical results of experiments

The numerical results comparing the accuracy of recognition in both forms of imagery using CNN and classical methods will be based on the multiple cross validation approach. The whole set of data in this method is split randomly into 2 parts. The learning data set is composed of 15 representatives of each class and the testing one on the rest (5 representatives of the class). The learning/testing runs have been repeated 10 times at random split of the data. In each repeated experiment the testing relative error was estimated. The final error is the average of all runs. The results will be limited to only testing cases, as the most representative.

The optimal CNN classification system was defined on the basis of pre-trained ALEXNET [5] after series of experiments with different number of neurons and their parameters in fully connected layers [5]. The final ALEXNET structure of CNN is shown in Fig. 4.

Image Input:227x227x3
Convolution1: 96 11x11x3 convolutions with stride [4 4] and zero-padding [0 0]
ReLU
Cross Channel Normalization with 5 channels per element
Max Pooling: 3x3 max pooling with stride [2 2] and zero-padding [0 0]
Convolution2: 256 5x5x48 convolutions with stride [1 1] and zero-padding [2 2]
ReLU
Cross Channel Normalization with 5 channels per element
Max Pooling: 3x3 max pooling with stride [2 2] and zero-padding [0 0]
Convolution3: 384 3x3x256 convolutions with stride [1 1] and zero-padding [1 1]
ReLU
Convolution4 :384 3x3x192 convolutions with stride [1 1] and zero-padding [1 1]
ReLU
Convolution5: 256 3x3x192 convolutions with stride [1 1] and zero-padding [1 1]
ReLU
Max Pooling: 3x3 max pooling with stride [2 2] and zero-padding [0 0]
Fully Connected Layer: 4096 elements of the vector fully connected to next layer
ReLU
Dropout: 50%
Fully Connected Layer: 2500 fully connected layer
ReLU
Dropout 50%
Fully Connected Layer: 50 fully connected neurons
Softmax classifier
Output layer: 50 neurons representing 50 classes

Fig. 4 The optimized CNN ALEXNET structure used in image recognition

It is composed of 5 convolution layers, composed of linear convolution filters followed by ReLU activation, normalization and max pooling. The fully connected layer starts from 4096 elements created from signals of the last convolution layer with application of ReLU function. The elements of this layer are connected to softmax classifier with the intermediate ReLU layer containing 2500 neurons and using the dropout coefficient equal 0.5. The softmax layer is composed of 50 neurons, representing 50 classes of images.

Thanks to using the predefined CNN ALEXNET the relatively small population of learning data base of images was enough to fit the final parameters of the network.

The description of the convolutional layers (for example: 96 11x11x3) includes the number of neurons in the layer (for example 96), the size of analyzing filter (in this example 11x11) and the number of images from the previous layer taking part in convolution (3 RGB input images in the first layer and the number chosen by user in the next layers, which is smaller or equal to the images in the preceding layer). The pooling operations in the layers applied the field 3x3, which means that only 1/9 part of input information has been preserved.

The typical learning curve in CNN training using Matlab [5] is presented in Fig. 5. It refers to training CNN structure on the basis of visible images. The mini batch learning accuracy and testing (validation) accuracy are plotted in the succeeding iterations.

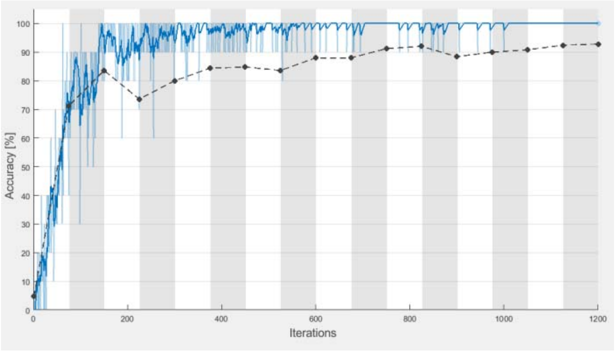

Fig. 5. The illustration of exemplary learning process of CNN for visible face images. The continuous (blue) line represents the learning and dash line the validation accuracy on mini batches.

In the case of classical approach the number of diagnostic features in each method was the same and equal 19. The SVM classifier has used Gaussian kernel of $\gamma=1$ and regularization coefficient $C=1000$.

The statistical results concerning accuracy in recognition of 50 classes of face images are presented in Table 1. They are given in the form of average misclassification rate and the standard deviation obtained in all cross validation experiments.

The results show that CNN is evidently the best. PCA (linear and nonlinear) and tSNE approaches belonged to the least efficient. This conclusion is true for both types of face imagery.

Table 1. The average misclassification rate (mean+/-std) of 50 classes of faces, committed by CNN and SVM supplied by different features generated by PCA, KPCA and tSNE.

	CNN [%]	PCA [%]	KPCA [%]	tSNE [%]
Visual images	4.42±1.35	13.30±1.5	12.96±1.4	15.84 ±1.6
Infra-red images	5.89+/-1.84	14.21+/-1.7	14.32±1.8	16.14±1.9

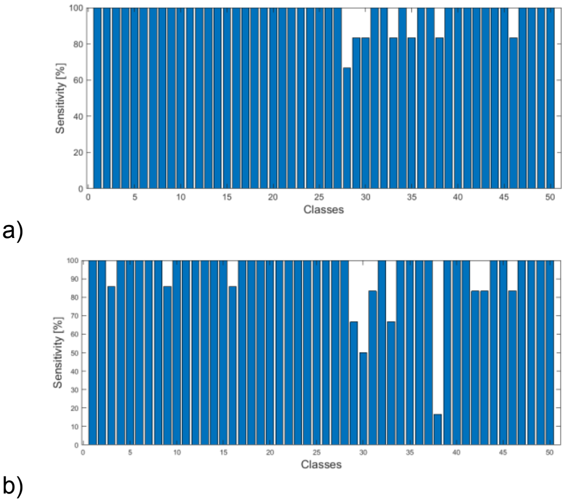

Fig. 6 Plot of sensitivity of CNN in recognition of the particular classes of faces: a) visible representation, b) IR representation.

The significant measure of system quality is also the sensitivity (the true recognized cases related to all cases representing particular class) of the classification system. Fig. 5 presents these results in a graphical form for the best approach (CNN) in one run of the system. The upper figure depicts the visible representation and bottom one – the IR one. The average sensitivity value was equal 97.33% for visible images and 94.23% for IR representation. The visible representation has allowed getting better results in terms of both, accuracy and sensitivity of face recognition.

Conclusions

The paper has shown the comparison of methods of face recognition applied to two types of images. The CNN approach does not need special image preprocessing. The originally acquired images are directly presented to the input side of the network and many hidden layers are responsible for simultaneous generation of diagnostic features and final recognition tasks. In classical approach to the problem the user is responsible for elaboration of special image descriptions and this stage is separated from the classification task.

The numerical experiments performed for recognition of 50 classes of faces have shown high advantage of CNN approach over the classical one, irrespective of the type of applied imagery of the face. The average misclassification rate is few times smaller than this obtained in the best classical approach to the image recognition. The visible representation of the faces was found better than the IR one.

Authors: *dr hab. inż. Krzysztof Siwek, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Email: ksiwek@iem.pw.edu.pl.*
prof. dr hab. inż. Stanisław Osowski, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Military University of Technology, Institute of Electronic Systems, Email: sto@iem.pw.edu.pl.

REFERENCES

[1] Belhumeur P., Hespanha J., Kriegman D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, *IEEE Trans. PAMI*, 19 (1997), no 7, 711–720
[2] Goodfellow I., Bengio Y., Courville A., Deep learning (2016), MIT Press, MA.
[3] Kong S., Heo J., Abidi B.R, Paik J. Abidi M. A., Recent advances in visual and infrared face recognition – a review, *Comput. Vision Image Understanding*, 97, (2005), 103-135
[4] Krizhevsky A., Sutskever I., Hinton G., Image net classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25*, (2012). 1-9.
[5] Matlab user manual – image toolbox, (2017b), MathWorks, Natick
[6] Schölkopf B., Smola A., Learning with kernels, (2002), Cambridge, MIT Press, MA
[7] Siwek K., Osowski S., Comparison of methods of feature generation for face recognition, *Przegląd Elektrotechniczny*, 90, (2014), nr. 4, 206-209
[8] Socolinsky D. A., Selinger A., Neuheisel J. D., Face recognition with visible and thermal infrared imagery, *Comput. Vision Image Understanding*, 91, (2003), 72-114
[9] Tan P. N., Steinbach M., Kumar V., Introduction to data mining, (2006), Pearson Education Inc., Boston.
[10] Van der Maaten L., Matlab toolbox for dimensionality reduction, v0.8.1, (2013), *Delft University of Technology*
[11] Wu S., Lin W., Xie S., Skin heat transfer model of facial termograms and its application in face recognition, *Pattern Recognition*, 2008, vol. 41, pp. 2718-2729
[12] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet.