# Bag of Features vs Deep Neural Networks for Face Recognition

Eliza Rebeca Tomodan
*Applied Electronics Department*
*Politehnica University Timișoara*
Timișoara, România
eliza.rebeca.tomodan@gmail.com

Cătălin Daniel Căleanu
*Applied Electronics Department*
*Politehnica University Timișoara*
Timișoara, România
catalin.caleanu@upt.ro

***Abstract*—This paper proposes a comparative study of Bag of Features (BoF) and Deep Neural Networks (DNN) approaches for the problem of face recognition. For the latter approach we consider three pre-trained models, namely AlexNet, ResNet50 and GoogleNet provided through Caffe Model Zoo and use them as feature extractors. Although these models were trained on different datasets, e.g., ImageNet, bottom-most layers act like universal feature extractors thus it is possible to be employed for different classification tasks. In order to adapt the models to various face datasets requirements we performed modifications to the input data as well as to the output layer of the pre-trained models by replacing it with a multiclass SVM classifier.**

***Keywords—Bag of features; Deep Convolutional Neural Networks;Face Recognition***

## I. INTRODUCTION

Face analysis represents today a very important biometric tool having multiple facets e.g. establishing identity, expression, age or gender for a certain person. The importance of the tackle research topic is justified trough the impressive areas of applications: market research, surveillance systems, automotive, medicine or human-machine interfaces and proved through numerous software libraries/APIs addressing human analytics: Fraunhofer's SHORE [1], Microsoft Face API [2] or Kairos' APIs & SDKs [3] as depicted in fig. 1.

The Bag of Features (BoF) principle introduced in [4] (referred as Bag of Word (BoW) within the context of document representation and Bag of Visual Words (BoVW) for image classification) was one of the most successful approaches for image classification, outperforming other feature learning algorithms [5]. Pit it versus cutting-edge technologies like Deep Convolutional Neural Networks and its associated learning paradigm - Deep Learning – in the face imagery would be of large interest for the scientific community.

The fundamental difference between the two paradigms is that in deep learning approach the features are automatically discovered, starting with low-level features up to high level features (fig. 2) whereas in traditional machine perception the features are hand tuned [6].


Multiple modules are processing a detected face in order to establish identity, expression, age, gender, etc. of the subject.

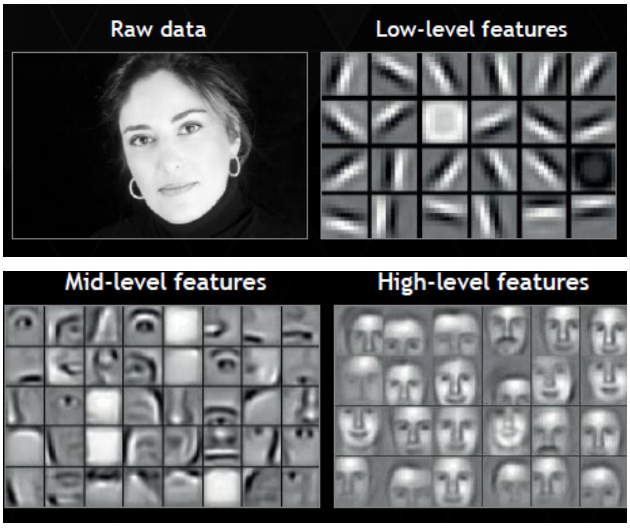


Fig. 2. Multiple modules are processing a detected face in order to establish identity, expression, age, gender, etc. of the subject.

978-1-5386-5925-0/18/$31.00 ©2018 European Union

## II. THE PARADIGMS

### A. Bag of Features

There are multiple BoF implementation variants mentioned in the literature. For example, in [7] are using randomly selected patches of image then perform a Zero Component Analysis followed by a K-means clustering for learning a dictionary in an unsupervised manner. Others are using BoVW in conjunction with Histogram of Oriented Gradients (HOG) [8]. In this case the image patches are given to a HOG descriptor for feature extraction and codebook calculation.

In the following we are formulating the general steps of the BoF paradigm and provide implementation details:

1) Detection and feature extraction/descriptor. Point selection is performed in our case using a fixed 4x4 pixels grid. In order to extract multi-scale features the descriptors are extracted from blocks of square 32, 64, 96, and 128 pixels size. The chosen feature point descriptor, typically in the form of Scale Invariant Feature Transform (SIFT) [9], is in our case Speeded up Robust Features (SURF) [10] as it has been shown to provide better results [11].

2) Create the codebook (dictionary or visual vocabulary). The output vectors from the feature extraction faze are clustered, in our case using K-mean clustering algorithm.

3) Quantize features using visual vocabulary. The above results are used to create a histogram of length *k*, the same *k* number as the number of generated clusters from K-mean. For our case k = 500 visual words.

4) Represent images by frequencies of "visual words". The algorithm creates a feature vector, a histogram of visual word occurrences in an image which represents a new reduced representation of an image. Encoded training images from each category are fed into a classifier in our case a multiclass linear SVM.

The whole process is illustrated by fig. 3.

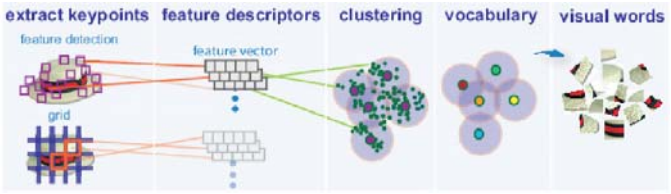


Fig. 3. The process of creating a bag of visual words.

### B. Deep Convolutional Neural Networks

Deep Neural Networks (DNN) are biologically inspired neural architectures, similar to some extent with the shallow perceptrons, having also particularities: high number of hidden layers implementing elementary operations like convolutions and spatial pooling using neurons with nonlinear activation - most of them of Rectified Linear Unit (ReLU) type - to compute activations of all convolved extracted features. Considering them as potential solutions for implementing soft biometrics tasks, e.g., face recognition is a fully motivated endeavor since they become a de fact standard in the field of Computer Vision.

As the main drawback of DNNs and theirs associated learning paradigm - Deep Learning - one could identify the very long training time even if clusters of GPUs are used as hardware support. An interesting concept comes to address this issue: transfer learning. As humans use the same visual system in order to perform different visual tasks it is possible to reuse some of the layers of a pretrained DNN as they act as universal feature extractors. It remains to retrain and adapt to a specific problem just the output/classification layers. In this way, the DNN technology might be available to a very broad range of classification tasks.

For the purpose of the present work we have chosen three DNN architectures, having a substantial variation in structure and size: AlexNet [12], GoogleNet [13] and ResNet50 [14]. A brief description of them is provided as follows.

AlexNet won ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 with only 15.3% error rate. The pretrained neural network AlexNet has 5 convolutional layers and 3 fully connected layers. AlexNet consists of 11x11, 5x5, 3x3 convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum (fig. 4).

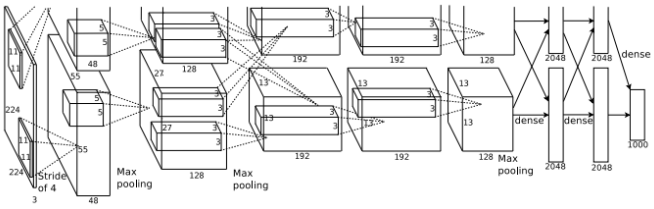


Fig. 4. AlexNet architecture [12].

The GoogleNet architecture (also known as Inception V1) won the ILSVRC competition in 2014. The error rate achieved was 6.67%, very close to the human level. This architecture uses 5x5, 3x3 and 1x1 convolutions and it consists of 22 layers, and 12 times less parameters than AlexNet, reducing the number from 60M to 4M, thus becoming much faster and more accurate than AlexNet. As seen in the fig. 5, GoogleNet replaced fully-connected layers at the end with simple global average pooling.

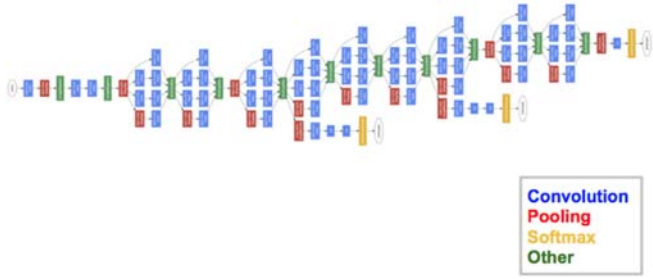


Fig. 5. GoogleNet/Inception V1 architecture [13].

The ILSVRC was won one year later, in 2015, with the score of 3.57%, using the framework of residual networks. The Residual Neural Network (ResNet) is a very deep NN, having 152 layers, that relies on micro-architecture modules (residual modules) as depicted in fig. 6.

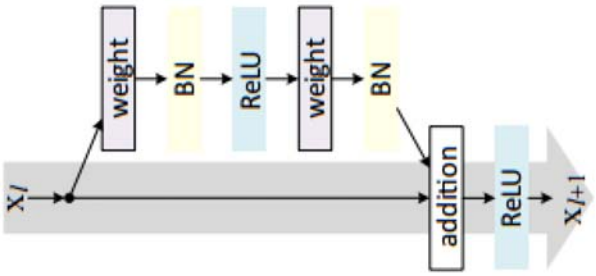


Fig. 6. The residual module in ResNet as proposed by [14].

In our experiments, for all three pre-trained networks, the fully connected (FC) layer was replaced by a multiclass SVM.

## III. Experimental Results

### A. The datasets

We are evaluating the above-mentioned paradigms against multiple custom and public face databases having important variations in terms of size, images/class, appearance, etc.

The first set of images are in-house acquired and contains a total of 45 images from 9 different peoples (fig. 7).

Fig. 7. Small in-house dataset.

The second database is represented by AT&T Laboratories Cambridge Database of Faces (fig. 8) with 10 different images of each of 40 distinct subjects thus having a total number of 400 images [15].

Fig. 8. A medium size dataset. Samples from AT&T Laboratories Cambridge Database of Faces.

The next sets are: faces94, faces95, faces96 and grimace provided by University of Essex [16].

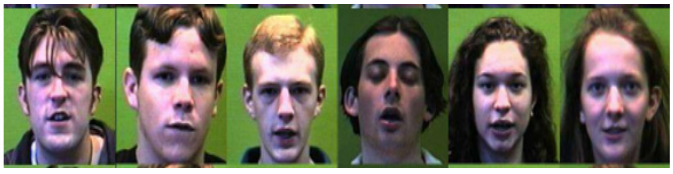

Fig. 9. Large dataset: ESSEX faces94 with a total of 3060 images of 153 people (20 images/person).

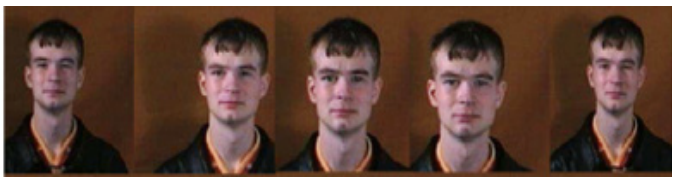

Fig. 10. Faces95 database is composed of 1440 images, of 72 individuals.

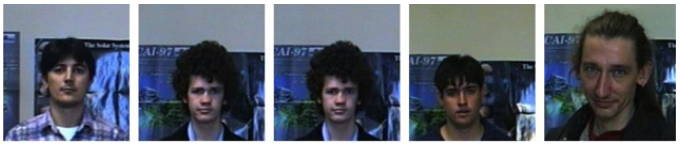

Fig. 11. Faces96 has 3040 images of 152 people.

Fig. 12. Grimace database has 360 images of 18 peoples.

### B. The test cases

For each experiment the available face images were split in half randomly into disjoint training and testing data. Then the reported accuracy was calculated as the average over 10 tests and the best and worst average results are highlighted.

In the first experiment we are studying the effect of the face database size over the classification accuracy provided by both BoF and DNN approaches. For this purpose, using images from Essex faces94, we generated subsequent databases of 40 images, 100 images, 500 images, 1000 images besides the original full-size dataset of 3060 images. The experimental results are provided in Table I.

TABLE I. The Effect of the Dataset Size

| Size | BoF | AlexNet | GoogleNet | ResNet50 |
|---|---|---|---|---|
| **40** | 100.00 | 100.00 | 100.00 | 100.00 |
| **100** | 100.00 | 100.00 | 100.00 | 100.00 |
| **500** | 97.8 | 98.44 | 95.64 | 98.96 |
| **1000** | 99.96 | 99.98 | 99.52 | 99.92 |
| **3060** | 99.63 | 99.49 | 98.81 | 99.30 |
| **Avg. acc.** | **99.45** | **99.58** | **98.79** | **99.63** |

The second experiment aims to study the effect of the number of images per class over the accuracy of the classification (see results in Table II).

TABLE II. Number of Images/Class

| Img/class | BoF | AlexNet | GoogleNet | ResNet50 |
|---|---|---|---|---|
| **5** | 100.00 | 100.00 | 100.00 | 100.00 |
| **10** | 97.05 | 96.70 | 88.35 | 95.70 |
| **20** | 99.66 | 99.32 | 98.85 | 99.59 |
| **Avg. acc.** | **98.90** | **98.67** | **95.73** | **98.43** |

The third test case focuses on the variation of the conditions in which the facial images are acquired thus using faces95, faces96 and grimace datasets. In this case the recognition is the most difficult for at least two different reasons:

- variation of background and scale

- extreme variation of expressions.

TABLE III. IMPORTANT VARIATION IN BACKGROUND AND EXPRESSION

| Dataset | BoF | AlexNet | GoogleNet | ResNet50 |
|---|---|---|---|---|
| **faces95** | 99.96 | 98.94 | 81.00 | 99.03 |
| **faces96** | 98.71 | 98.41 | 98.54 | 98.78 |
| **grimace** | 99.83 | 100.00 | 99.16 | 100.00 |
| **Avg. acc.** | **99.50** | **99.12** | **92.90** | **99.27** |

## IV. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Within the framework of face recognition, the current paper contributions are the comparative evaluation of the performances of two most important paradigms, namely BoF and DNN. For the latter, only the feature extraction layers where kept as they were formed initially for the ImageNet dataset. Next, we use the DNN image features to train a multiclass SVM classifier using a fast-stochastic gradient descent solver. This helps speed-up the training when working with high-dimensional DNN feature vectors.

We have also been able to demonstrate that the our BoF implementation, using SURF method as point detector/extractor outperforms on average the pre-trained models (see Table IV).

TABLE IV. OVERALL ACCURACY FOR ALL 3 TEST CASES [%]

| BoF | AlexNet | GoogleNet | ResNet50 |
|---|---|---|---|
| 99.29 | 99.12 | 95.81 | 99.11 |

This contradict in some respects the findings of [17]. There are two possible explanations for this fact:

- the principle behind theirs BoF algorithm differs (in [17] a local/HOG-BOW type is employed)

- the database Wild-Anim used in [17] differs in substance and proprieties (5000 images, 5 classes)

As future research directions we would like to investigate:

- the problem of very large face recognition datasets

- train from the scratch the above mentioned DNN models

- design custom DNN which could exploit the particularities present in face imagery

## REFERENCES


[1] Fraunhofer, Face Detection Software SHORE®, https://www.iis.fraunhofer.de/en/ff/sse/ils/tech/shore-facedetection.html

[2] Mirosoft Face API, https://azure.microsoft.com/en-us/services/cognitive-services/face/

[3] Kairos' APIs & SDKs, https://www.kairos.com/

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in Computer Vision (ECCV), 8th European Conference on, 2004, pp. 1–22.

[5] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in International conference on artificial intelligence and statistics, 2011, pp. 215–223.

[6] L. Brown, "Deep Learning with GPUs, GEOINT2015, http://www.nvidia.com/content/events/geoInt2015/LBrown_DL.pdf

[7] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2012, pp. 1098–1105.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society Conference on, vol. 1, 2005, pp. 886–893.

[9] D. Lowe, "Object recognition from local scaleinvariant features", proceedings of 7th International Conference on Computer Vision, (2), 1999, pp. 1150-1157.

[10] H. Bay, T. Tuytelaars and L. Gool, "SURF: Speeded Up Robuts Features", Computer Vision ECCV, 3951, 2006, pp. 404-417.

[11] K. Ahmad et. al, "Evaluation of SIFT and SURF Using Bag of Words Model on a Large Dataset", Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (3) 2013, pp. 492-495.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[13] C. Szegedy et. al, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", https://arxiv.org/abs/1512.03385, 2015.

[15] AT&T Database of Faces, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[16] Collection of Facial Images, University of Essex, http://cswww.essex.ac.uk/mv/allfaces/index.html

[17] E. Okafor et al., "Comparative study between deep learning and bag of visual words for wild-animal recognition," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-8. doi: 10.1109/SSCI.2016.7850111.