Feature-level Fusion vs. Score-level Fusion for Image Retrieval Based on Pre-trained Deep Neural Networks

Nikolay Neshov*, Krasmir Tonchev, Agata Manolova, Vladimir Poulkov and Georgi Balabanov

Faculty of Telecommunications, Technical University of Sofia, bul. Kl. Ohridski 8, Sofia 1000, Bulgaria
E-mail: nneshov@tu-sofia.bg; k_tonchev@tu-sofia.bg; amanolova@tu-sofia.bg; vkp@tu-sofia.bg; grb@tu-sofia.bg
**Corresponding Author*

Received 09 October 2023; Accepted 18 February 2024;

Abstract

Today's complex multimedia content made retrieving images similar to the user's query from the database a challenging task. The performance of a Content-Based Image Retrieval System (CBIR) system highly depends on the image representation in a form of low-level features and similarity measurement. The traditional visual descriptors that do not provide good prior domain knowledge could lead to poor performance retrieval results. On the other hand, Deep Convolutional Neural Networks (DCNNs) have recently achieved a remarkable success as methods for image classification in various domains. Recently, pre-trained deep convolution neural networks on thousands of classes have the ability to extract very accurate and representative features which, in addition to classification, can also be successfully used in image retrieval systems. ResNet152, GoogLeNet and InceptionV3 are some of the effective and successful examples of pre-trained DCNNs recently applied in a computer vision tasks such as object recognition, clustering, and classification. In this paper, two approaches for a CBIR system, namely

Journal of Mobile Multimedia, Vol. 20_4, 769–784.
doi: 10.13052/jmm1550-4646.2041

early fusion and late fusion, have been presented and compared. The early fusion utilizes concatenation of the features extracted by each possible pair of DCNNs, that is ResNet152-GoogLeNet, ResNet152-InceptionV3, and GoogLeNet-InceptionV3, and the late fusion apply CombSum method with Z-Score standardization to combine the score results provided by each DCNN of the aforementioned pairs. In the experiments on a popular WANG dataset it has been shown that late fusion approach slightly outperforms early fusion approach. The best performance of our experiments in terms of Average Precision (AP) for the top 20 results reaches 96.82%.

Keywords: Content-based image retrieval system (CBIR), late fusion, early fusion, deep convolutional neural networks (DCNNs), ResNet152, GoogLeNet, InceptionV3.

1 Introduction

In recent years, the exponential growth of digital images on the Internet has made the process of searching for similar images increasingly complex, leading to a surge of interest in computer vision research. Traditional approaches to image retrieval, relying on global or local descriptors, often struggle to perform well on large and heterogeneous image datasets [1]. However, the emergence of deep learning has revolutionized image classification tasks, with various significant Deep Convolutional Neural Network (DCNN) models developed, including VGG, GoogLeNet, AlexNet, ResNet, and InceptionV3. These pre-trained DCNNs, leveraging datasets like ImageNet, exhibit high feature encoding capabilities that can be effectively utilized in Content-Based Image Retrieval (CBIR) systems, supplanting traditional descriptors. For instance, Ahmed A. [2] introduced a CBIR system utilizing pre-trained CNN models ResNet18 and SqueezeNet to extract two groups of features for online image retrieval. Alzu'bi et al. [3] proposed the CRB-CNN, a bilinear CBIR system modeling two DCNNs in parallel, resulting in highly discriminative and compact features extracted from VGG-16 and VGG-m. Nguyen et al. [4] developed a deep network architecture for classifying microscopic cell images using transfer learning, concatenating features from three pre-trained DCNNs in the learning phase. Similarly, Rajkumar and Sudhamani [5] proposed a CBIR system based on residual neural networks, computing image similarity using Euclidean distance between feature vectors. Kumar S., et al. [6] combined information from DarkNet-19 and DarkNet-53 in their CBIR system for image retrieval. Ouhda et al. [7]

Table 1 Summary of related works

Authors	Approach	Key Results
Ahmed A. [2]	Utilized ResNet18 and SqueezeNet	Efficient online image retrieval
Alzu'bi et al. [3]	Developed CRB-CNN with VGG-16 and VGG-m	Highly discriminative features
Nguyen et al. [4]	Used transfer learning with 3 DCNNs	Effective classification of cells
Rajkumar and Sudhamani [5]	Leveraged residual neural network	Accurate image similarity computation
Kumar S., et al. [6]	Combined information from DarkNet-19 and DarkNet-53	Successful image retrieval
Ouhda et al. [7]	Combined DCNNs with SVM for CBIR tasks	Enhanced efficiency in image retrieval
Ahmed and Mohamed [8]	Employed both early and late fusion approaches	Improved fusion strategies for CBIR

designed an approach that combines DCNNs with Support Vector Machine (SVM) for efficient CBIR tasks. Additionally, Ahmed and Mohamed [8] applied both early and late fusion approaches, combining color and texture features in the early fusion phase and employing common distance measures in the late fusion phase. However, none of the aforementioned works have examined combination of features vs. combination of scores that is the main goal of our work. The overview of the previously mentioned related studies is presented in Table 1.

2 Image Database

The database Wang [9] used in our experiments is made up of 1000 color images divided into 10 classes of 100 images each. This balanced distribution enables unbiased evaluation of algorithms across different categories.

Every group contains 265×384 or 384×256 pixel resolution images. The 10 classes of this database are: African people, Bus, Dinosaur, Flower, Beach, Elephant, Buildings, Food, Horse, and Mountain (Figure 1). These categories include natural scenes, objects, people, animals, landscapes, and more. Researchers extensively utilized this database for testing various features due to its abundant class information, which facilitated robust evaluation and comparison of algorithms. The WANG database resembles typical stock photo retrieval scenarios, where users may seek images similar to their own but with specific preferences, such as lower royalties or uniqueness. Its 10

Figure 1 Some samples from Wang database.

Figure 2 The basic structure of Resnet152.

classes serve as benchmarks for relevance estimation: when a user provides a query image, it's presumed they desire images from the same class. As such, the other 99 images within the same class are deemed relevant, while images from different classes are deemed irrelevant.

3 Pre-trained Deep Convolutional Neural Networks

In this section, the three convolutional neural networks (ResNet152, GoogLeNet, and InceptionV3) utilized as a base for fusion have been briefly described. All possible pairs of networks for combination were constructed and compared, examining Feature-Level fusion versus Score-Level fusion, as detailed in Section 4.

3.1 ResNet152

Residual Network (ResNet) [10] is a deep learning model that solves the problem of the vanishing/exploding gradient by introducing a structure called skip connections. This structure connects activations of a layer to later layers by skipping some layers in between. This forms residual unit. The ResNet is formed by stacking several residual blocks together. For the experiments in our work, ResNet152 was chosen because is achieves the best accuracy among all Resnet family members [10]. On Figure 2 is illustrated the basic structure of ResNet152.

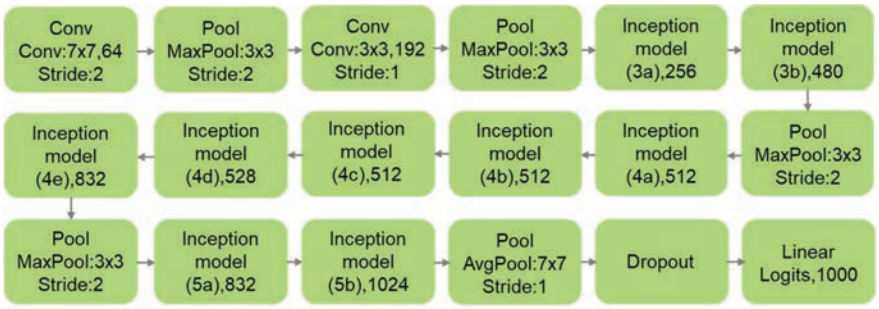


Figure 3 The basic structure of GoogLeNet.

3.2 GoogLeNet

The GoogLeNet model, or InceptionV1, was introduced by Szegedy et al. [11]. It's architecture is based on the "Inception module", which have convolution layers in parallel and this allows the model to choose between multiple convolutional filter sizes in each block. GoogLeNet is 22–layer network and has the advantage of more efficient computational requirements compared to other networks that have similar depths. The basic structure of GoogLeNet is depicted in Figure 3.

3.3 InceptionV3

InceptionV3 [12] is an extension of GoogLeNet [11]. The main innovation in InceptionV3 is better model adaption by using multiple strategies to reduce the number of model parameters. This includes: spatial factorization into asymmetric convolutions, utilizing auxiliary classifiers, factorization into smaller convolutions, and efficient reduction of grid size. Compared to InceptionV1, InceptionV3 reduces the amount of computations and improves efficiency without loss of speed. The basic structure of InceptionV3 is shown in Figure 4.

4 Algorithm Description

In this section, the structures of the experimented algorithms are explained. Initially, the traditional algorithm of a CBIR system based on a single pre-trained DCNN is described, followed by an overview of the two adopted methods for fusion. These are the feature, and the score level fusion utilizing two DCNNs.

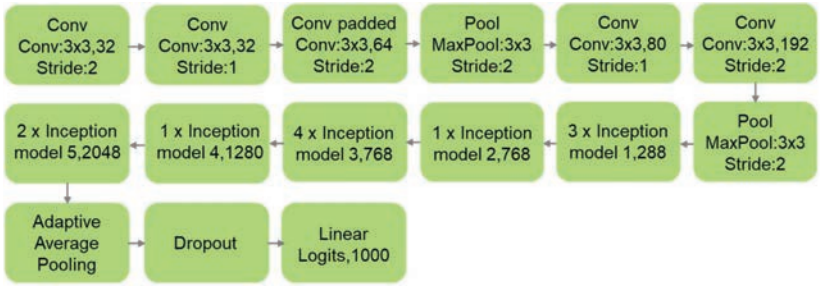


Figure 4 The basic structure of InceptionV3.

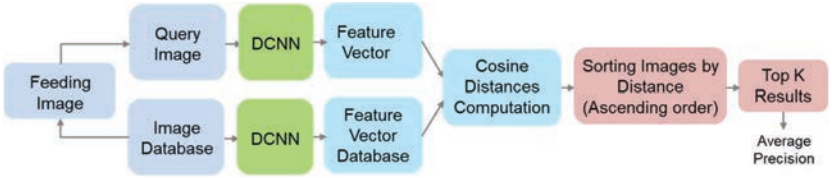


Figure 5 Traditional CBIR system using single DCNN.

4.1 Image Retrieval Using Single Pre-trained DCNN

On Figure 5 is shown the architecture and the assessment process of traditional CBIR system, using DCNN. The images from the database are processed by the DCNN and feature vectors are extracted from the latest layer before fully connected layer. Thus, a feature vector database is formed. For the assessment purpose, each image from the database is fed as a query image to the system and the corresponding feature vector is computed by DCNN.

Further the distances between the query feature vector and the feature vectors from the database are computed (excluding the feature vector of the query). The experiments incorporated the cosine distance as the chosen metric for assessing the similarity between feature vectors. The images from the database are sorted is ascending order with respect to the distances and the top K images are returned. The precision for the given query is then computed. The process is repeated for each query (each image from the database) and the average precision at the top K results ($p@K$) across all the queries is calculated. Precision is chosen as it measures the relevance of retrieved images, crucial for user satisfaction in CBIR. Evaluating precision at the top K results reflects real-world user behavior and algorithm effectiveness. Maximizing precision enhances the quality of retrieved images, aligning with system goals and user expectations.

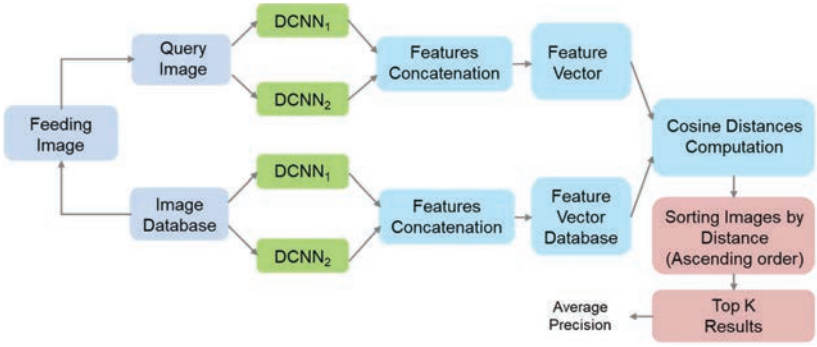


Figure 6 CBIR system utilizing feature concatenation (early fusion) of two DCNNs.

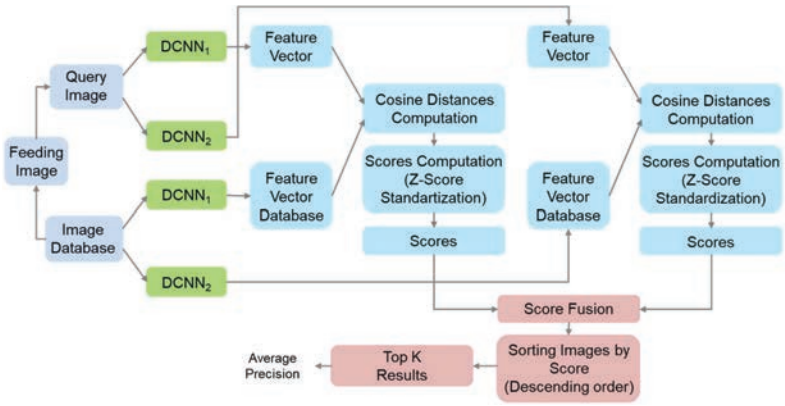


Figure 7 CBIR system using score combination (late fusion) of two DCNNs.

4.2 Feature-Level Fusion (Early Fusion)

On Figure 6 is depicted a diagram of architecture of CBIR system that utilizes two DCNNs to form feature vector for a given image by concatenation of the corresponding feature's components. The operation principle of this architecture is similar to that described in Section 4.1.

4.3 Score-Level Fusion (Late Fusion)

On Figure 7 is drown the diagram of the proposed architecture of CBIR system that combines the image scores resulted by each DCNN.

This architecture shares the same assessment scenario as that described in Section 4.1 as well as the same processing of images for distances computation by both DCNNs. That is, it consists of two single traditional CBIR

Table 2 Precision at the top 20 results for the Wang [9] database reached by each DCNN individually and by the two fusion approaches (early and late) utilizing DCNN pairs

Deep Convolutional Neural Network(s)	*p@20,%*
ResNet152	96.00
GoogLeNet	94.91
InceptionV3	95.66
ResNet152 – GoogLeNet (Early fusion)	96.12
GoogLeNet – InceptionV3 (Early fusion)	96.25
ResNet152 – InceptionV3 (Early fusion)	96.79
ResNet152 – GoogLeNet (Late fusion)	96.42
GoogLeNet – InceptionV3 (Late fusion)	96.44
ResNet152 – InceptionV3 (Late fusion)	**96.82**

systems that produce two unsorted lists of distances (one for each network) for the corresponding database feature vectors to the query's feature vector. Having both unsorted list of N features/images, the formation of final list of results can be described as follows. Let $D^1 = \{d_1^1, \ldots, d_n^1, \ldots, d_N^1\}$ and $D^2 = \{d_1^2, \ldots, d_n^2, \ldots, d_N^2\}$ are the lists of distances of the 1-st and 2-nd DCNN correspondingly. d_n^1 is the distance between a query feature vector and *n*-th image's feature vector in the database for the first DNN, and d_n^2 – for the second DNN. Both lists of distances are transformed to list of scores of similarity $S^1 = \{s_1^1, \ldots, s_n^1, \ldots, s_N^1\}$ and $S^2 = \{s_1^2, \ldots, s_n^2, \ldots, s_N^2\}$, where s_n^1 and s_n^2 are computed by:

$$s_n^m = d_{max}^m - d_n^m, \tag{1}$$

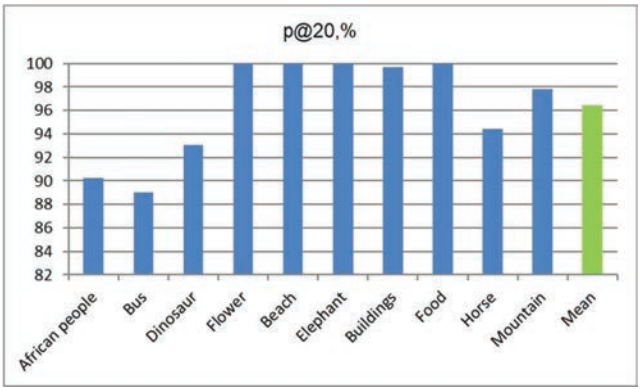
where m is the m-th DNN ($m = 1, 2$), and $d_{max}^m = \max_{n=1\ldots N}\{d_n^m\}$.

It should be noted that the range of similarity scores for each list may be different. Thus, differences in the dimensionality of individual features affect the value of s_n^m, which is undesirable. This necessitates the need for normalization before calculating the final value. Z-Score standardization was used for this purpose. The normalized score $\overline{s}_n^m$ is computed as:

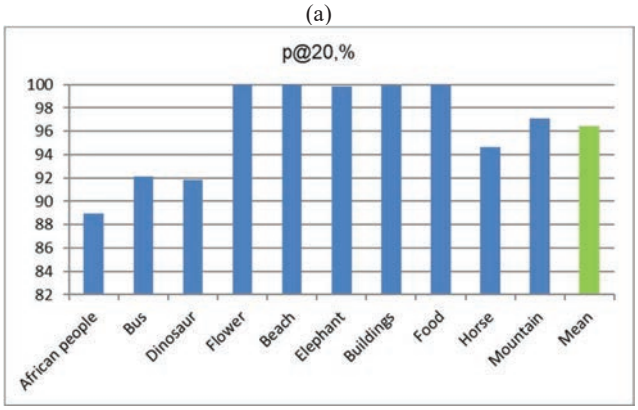
$$\overline{s}_n^m = \frac{s_n^m - \mu^n}{\sigma^m}, \tag{2}$$

where μ^n is the mean and σ^m is the standard deviation of the series of similarity scores in S^m. The fused score $\overline{s}_n^f$ for the n-th image is given by:

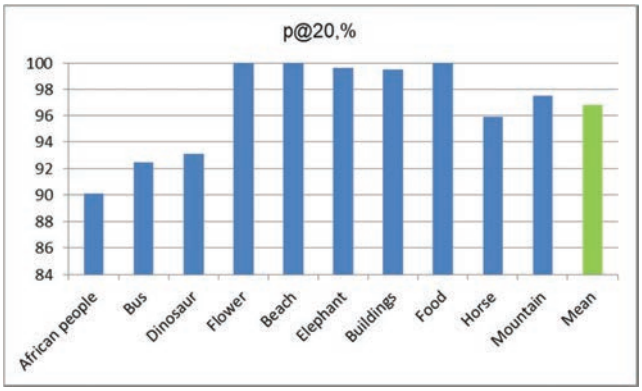
$$\overline{s}_n^f = \frac{1}{M}\sum_{m=1}^{M} \overline{s}_n^m. \tag{3}$$



(a)



(b)



(c)

Figure 8 Precision at the top 20 results for the Wang [9] dataset by categories utilizing late fusion approach for the ResNet152 – InceptionV3 (a), GoogLeNet – InceptionV3 (b), and ResNet152 – InceptionV3 (c) pairs.

The list of fused normalized scores can be written as: $S^f = \{\overline{s}_1^f, \ldots, \overline{s}_n^f, \ldots, \overline{s}_N^f\}$. This list is further sorted in descending order with respect of the normalized scores and the corresponding top K images are given as a result.

5 Experimental Results

In Table 2 is shown the performance results in terms of precision at the top 20 results reached by each DCNN individually and by the two types of fusion between the different DCNN pairs. It can be seen that each fusion scheme gives better performance than all single neural network realization. It also can be seen that the value of $p@20$ is better for the late fusion approaches compared to that of early fusions. The highest difference is 0.3% for the ResNet152 – GoogLeNet pair. The best performance pair is ResNet152 – InceptionV3 utilizing late fusion with $p@20$=96.82%. The corresponding early fusion approach for the same pair reaches nearly the same performance as $p@20$=96.79%, however still a bit lower.

Figure 8a, b, and c depicts the distribution of precision of the top 20 results for the Wang [9] database by categories for the late fusion schemes of ResNet152 – GoogLeNet, GoogLeNet – InceptionV3, and ResNet152 – InceptionV3 pairs respectively. For the most distinguishable categories Flowers, Beaches, and Foods, it can be seen that the $p@20$ reaches almost 100%, while for the Africans it is near 90% which is the worst case. The precision for the Dinosaur, Horse, and Mountain categories stabilizes in the middle ground between approximately 92% and 98%.

6 Conclusion

In this article, a comparison of CBIR systems utilizing DCNNs is presented. A CBIR system employing an enhanced late fusion scheme was built, demonstrating superior performance compared to experiments utilizing single neural networks and those employing feature concatenation. The increase in $p@10$ is relatively small (0.3%), but the important thing is that the late fusion has a potential to beat early fusion. However, most of the recent works are concentrated on the early fusion methods, where combining of features gets higher accuracy/precision. Our key contribution is that late fusion is a method that future research should consider. Despite the promising potential highlighted in our study, late fusion approaches remain underexplored in

the current CBIR research landscape. Recent works predominantly focus on early fusion methods, attributing higher accuracy and precision to early-stage feature combination. This limited attention to late fusion underscores the need for more exploration and consideration in future research initiatives within the field. Our future work will prioritize the exploration of specific late fusion techniques like voting schemes, attention mechanisms, and ensemble methods to leverage the strengths of individual networks. Additionally, integrating domain knowledge and developing novel fusion strategies adaptable to diverse datasets and scenarios will be crucial areas of focus. By embracing these recommendations, we aspire to drive innovation in content-based image retrieval.

Acknowledgements

This research is financed by the European Union-Next Generation EU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0005: "Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia" (IDEAS).

References

[1] A.C. Valente, F.V.M. Perez, G.A.S. Megeto, M.H. Cascone, O. Gomes, T.S. Paula and Q. Lin, "Comparison of texture retrieval techniques using deep convolutional features", in *Proc. IS&T Int'l. Symp. on Electronic Imaging: Imaging and Multimedia Analytics in a Web and Mobile World*, 8, pp. 406-1–406-7, 2019.

[2] A. Ahmed, Pre-trained CNNs Models for Content based Image Retrieval, *International Journal of Advanced Computer Science and Applications*, 2021.

[3] A. Alzu'bi, A. Amira and N. Ramzan, "Content-Based Image Retrieval with Compact Deep Convolutional Features," *Neurocomputing*, 249, pp. 95–105, 2017.

[4] L.D. Nguyen, D. Lin, Z. Lin and J. Cao, "Deep CNNs for Microscopic Image Classification by Exploiting Transfer Learning and Feature Concatenation," *Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, Seville, Spain, pp. 1–5, May, 2018.

[5] R. Rajkumar and M.V. Sudhamani, "Image Retrieval System Using Residual Neural Network in a Distributed Environment," *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp. 2277–3878, 2020.

[6] S. Kumar, M.K. Singh and M. K. Mishra, "Improved Content-Based Image Retrieval Using Deep Learning Model," *Journal of Physics: Conference Series*, Ser. 2327 012028, 2022.

[7] O. Mohamed, E.A. Khalid, O. Mohammed and A. Brahim, "Content-Based Image Retrieval Using Convolutional Neural Networks," In *First International Conference on Real-Time Intelligent Systems*, Springer, Cham, Switzerland, pp. 463–476, 2019.

[8] A. Ahmed and S. Mohamedc, "Implementation of Early and Late Fusion Methods for Content-Based Image Retrieval," *Int. J. Adv. Appl. Sci.* vol. 8, no. 7, pp. 97–105, 2022.

[9] J.Z. Wang, J. Li and G. Wiederhold, "Simplicity: Semantics-Sensitive Integrated Matching for Picture Libraries," *IEEE Trans. Pattern Anal. Mach. Intell.* Vol. 23(9), pp. 947–963, 2001.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770–778, 2016.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1–9, 2015.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 2818–2826, 2016.

Biographies

Nikolay Neshov holds a PhD in Communication and Computer Technology from Technical University of Sofia. His doctoral research was concentrated in optimization of Content-Based Image Retrieval (CBIR) systems based on probabilistic models. Currently, his research interest covers computer vision, machine learning, decision analysis, video and image indexing and retrieval, text mining, and facial analysis.

Krasmir Tonchev a senior researcher leading research activities at the "Tele-Infrastructure Lab", Faculty of Telecommunications, TU-Sofa. His research interests include, on the theoretical side, large scale Kernel Machines, modelling of dynamical behaviour, Bayesian modelling, and on the application side, 2D and 3D facial analysis for soft biometrics, afective computing and general scene understanding from video. He is an IEEE member.

Agata Manolova is associate professor with the Faculty of Telecommunications at the Technical University of Sofa (TU-Sofa), Bulgaria and the head of the research laboratory "Electronic systems for visual information". Her domains of interest are machine learning, pattern recognition, computer vision, image and video processing, biometrics, augmented and virtual reality. She has received her PhD from Universite de Grenoble, France. She is laureate of Fulbright scholarship and an IEEE member.

Vladimir Poulkov is a full time professor at the Faculty of Telecommunications at TU-Sofia. His expertise is in the field of information transmission theory, modulation and coding interference suppression, power control and resource management for next generation telecommunications net-works, cyber physical systems. Currently he is Head of "Te-leInfrastructure" and "Electromagnetic Compatibility of Communication Systems" R&D Laboratories, chairman of Bulgarian Cluster Telecommunications, Vice-Chairman of European Telecommunications Standardization Institute (ETSI) General Assembly, Senior IEEE Member.

Georgi Balabanov is associate professor with the Faculty of Telecommunications at the Technical University of Sofia (TU-Sofia), Bulgaria. He received his PhD degree in Communication Networks and Systems from TU-Sofia. He is an affiliate researcher at TeleInfrastructure R&D Lab. Dr. Balabanov has participated in several scientific projects – both national and international. His research interests include Embedded Systems, Teletraffic Engineering, QoS, Internet of Things, Ambient Assisting Living Systems. He is an IEEE member.