Content Based Image Retrieval by Convolutional Neural Networks

Safa Hamreras[1], Rafaela Benítez-Rochel[2], Bachir Boucheham[1], Miguel A. Molina-Cabello[2], and Ezequiel López-Rubio[2]

[1] Department of Computer Science
University of 20 August 1955, BP 26, Route El Hadaiek, 21000 Skikda, Algeria
safahamreras@gmail.com,bachir.boucheham@hotmail.com,
[2] Department of Computer Languages and Computer Science. University of Málaga.
Bulevar Louis Pasteur, 35. 29071 Málaga. Spain.
{benitez,miguelangel,ezeqlr}@lcc.uma.es,
WWW home page: http://www.lcc.uma.es/~ezeqlr/index-en.html

Abstract. In this paper, we present a Convolutional Neural Network (CNN) for feature extraction in Content Based Image Retrieval (CBIR). The proposed CNN aims at reducing the semantic gap between low-level and high-level features. Thus, improving retrieval results. Our CNN is the result of a transfer learning technique using Alexnet pretrained network. It learns how to extract representative features from a learning database and then uses this knowledge in query feature extraction. Experimentations performed on Wang (Corel 1K) database show a significant improvement in terms of precision over the state of the art classic approaches.

Keywords: Content Based Image Retrieval, Convolutional Neural Networks, Feature extraction.

1 Introduction

The increased use of digital computers, multimedia, and storage systems over recent years has result in large image and multimedia content repositories. This huge amount of multimedia data is being used in many fields like medical treatment, satellite data, electronic games, archaeology, video and still images repository, and digital forensics and surveillance systems. That rapid growing has created an ongoing demand of retrieval images systems operating on a large scale.

Content Based Image Retrieval (CBIR) is the procedure of automatically retrieving images by the extraction of their low-level visual features, like color, texture, shape properties or any other features being derived from the image itself [27]. The performance of a CBIR system mainly depends on these selected features [27]. Thus, it can be said that through navigation, browsing, query-by-example etc, we can calculate the similarity between the low-level image contents which can be used for the retrieval of relevant images. The most challenging issue associated with CBIR systems is reducing the semantic gap. It is the information lost by representing an image in terms of its features i.e., from high level semantics to low level features [15]. This gap exists between the visual information captured by the imaging device and the visual information perceived by the human vision system (HVS) and it can be reduced either by embedding domain specific knowledge or by using some machine learning technique to develop intelligent systems that can be trained to act like HVS.

There has been a significant growth in machine learning research but mainly deep learning has already demonstrated its potential in large-scale visual recognition [2], [6], [16].The main reasons behind its success are the availability of large annotated data sets, and the GPUs computational power and affordability.

Deep learning is a subset of machine learning which uses a hierarchical level of artificial neural networks to carry out the process of machine learning. The term Deep Neural Network (DNN) refers to describe any network that has more than three layers of non-linear information stages in its architecture and Deep Learning (DL) is a collection of algorithms for learning in Deep Neural Networks, used to model high-level abstractions in data [28]. Thus, deep learning techniques gives a direct way to get feature representations by allowing the system (deep network) to learn complex features from raw images without using hand crafted features [12].

Deep learning has been successfully applied to many problems e.g., computer vision and pattern recognition [29], [25], [11], [20],[4],[5], computer games, robots and self-driving cars [18], [10], [23], [1], voice recognition and generation [21], [9], music composition [8], [3] and natural language processing [22].

Due the success of deep leaning, and the importance of feature extraction in CBIR systems, in this work we propose a CNN for learning feature representation in CBIR. The proposed CNN learns how to extract relevant features from a given images database and then applies this information in image retrieval process. The rest of this paper is organized as follows: Section 2 reviews the state of the art approaches in CBIR, Section 3 presents the adopted methodology, Section 4 reports the obtained results, and Section 5 concludes this paper.

2 State of the art

The used features and similarity measure are the two critical choices to make when building a CBIR system. Therefore most researches in CBIR focus on these topics to enhance these systems. In this section we discuss the major contributions that treat these points as well as a recent approach adopted in CBIR which is Deep Learning.

In feature representation level, recent traditional approaches focus on efficient representation of images by improving visual descriptors and combining them in order to improve retrieval results. Ekta Walia *et al.* [7] proposed a CBIR framework where they used late fusion techniques in order to improve the accuracy, the techniques used were: Borda Count, Min-Max normalization and Z-Score normalization. The fusion was performed on two descriptors: Modified

Color Difference Histogram (CDH) which was improved by using filtering on lab color space images and modified edge orientation, and The Angular Radial Transform (ART). Jing-Ming Guo *et al.* [13] designed a framework that generates an efficient feature vector using low complexity-Dither Block Truncation Coding (ODBTC). The result feature vector is composed of color co-occcurence feature (CCF) and bit pattern features (BPF). Similar images are then sorted based on the relative distance measure between query image and all database images. In another similar work, Jing-Ming Guo *et al.* [14] used Dot-Diffused Block Truncation Coding Features to create a feature vector composed of Color Histogram Feature (CHF) and Bit Pattern Feature (BPF). For similarity measure they used: L1 distance, x2 distance, Fu distance, and Modified Canberra Distance. Y.R. Charles *et al.* [26] used Local Mesh Texture Color Pattern descriptor to represent an image. In this work, there is a merge of three color spaces: l1l2l3, YIQ, YcbCr, where the three components: l1, Cb, Q are extracted from these color spaces in order to create three opponent texture patterns: l1Cb, CbQ, and l1Q. Finally, the three opponent patterns are fused in order to create one feature vector. Moreover they used four distance functions for similarity measurement: Manhattan, Euclidian, Canberra, and d1 distance measure. Xiang-Yang Wang *et al.* [24] proposed a system that combines color and texture features including: Pseudo-Zernike chromacity distribution moments in opponent chromacity space for color representation, and rotation-invariant and scale-invariant descriptor in steerable pyramid domain for texture representation.

Some other recent approaches are interested in similarity measurement. For example, ELALMI [19] proposed a model for CBIR where he injected a matching strategy to measure similarity between images, the proposed model extracts color and texture features from images, and then reduces this set by selecting relevant and non-redundant features. After that, ANN network is used to classify images so that the retrieved images are from the same class as the query image. For retrieval step, the model uses a matching strategy by calculating the area between image query features vectors and all database images features vectors.

Recently, researchers are using CNNs to learn image representation and similarity measure. One of the most interesting contributions belongs to Kevin Lin *et al.* [17]. The authors proposed a framework for image retrieval using CNNs where they use binary hash codes for less time consuming. Moreover, the binary hash codes are learned simultaneously while fine-tuning the network. For similarity measure, they first use Hamming distance in order to compare query image binary hash code and database images binary hash codes. The result set of similar images is then filtered and sorted so the most k similar images are extracted, for that they use the euclidean distance to measure similarity between features extracted from F7 layer.

In this paper, we focus on image representation in CBIR by learning efficient feature representation. In order to achieve this we use a CNN to carry out feature extraction process. The used method is described in the following sections.

3 Methodology

In order to design a Content Based Image Retrieval (CBIR) system based on a Convolutional Neural Network (CNN), we propose to employ a neural architecture which has an output layer with one output neuron for each object class in the training image database. Let D be the number of object classes. The CNN has to be trained on the images of the training database, where the desired output for each image is a unit vector of size C with a one at the component associated to the class of the object which is depicted in the training image, and zeros at the rest of the components. Given this configuration, the CNN learns to estimate the probabilities that an image represents an object class:

$$f\left(\mathbf{X}\right) = \left(P\left(C_1|\mathbf{X}\right), ..., P\left(C_D|\mathbf{X}\right)\right) \in \left[0,1\right]^D \tag{1}$$

where $\mathbf{X}$ is an input image, and C_i is the i-th object class, with $i \in \{1, ..., D\}$.

After the CNN is trained in this way, the system is ready to accept user queries. Let us note $\mathbf{X}_{Query}$ the query image, and $\mathbf{X}_j$ the training database images, where $j \in \{1, ..., N\}$ and N is the number of images in the database. Then the database images are ranked according to the Euclidean distances between the probability vector $f\left(\mathbf{X}_{Query}\right)$ associated to the query image, and the probability vectors $f\left(\mathbf{X}_j\right)$ associated to the database images. For example, the most similar image in the database can be obtained as follows:

$$s = \arg \min_{j \in \{1,...,N\}} \left\| f\left(\mathbf{X}_{Query}\right) - f\left(\mathbf{X}_j\right) \right\| \tag{2}$$

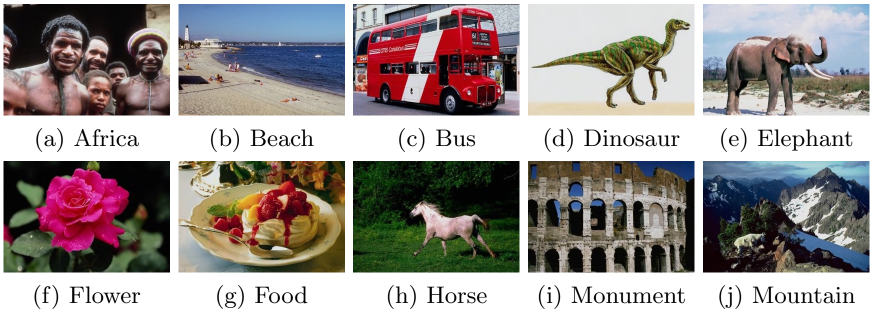
It is then possible to obtain the k most similar images in the database, as ranked by $\left\| f\left(\mathbf{X}_{Query}\right) - f\left(\mathbf{X}_j\right) \right\|$. Alternatively, a similarity threshold τ can be defined so that all images in the database with $\left\| f\left(\mathbf{X}_{Query}\right) - f\left(\mathbf{X}_j\right) \right\| < \tau$ are declared to be similar to the query image.

4 Experimental Results

In this section, we report the obtained results by our approach, which uses Convolutional Neural Networks, and compare them with other state of the art traditional approaches.

4.1 Methods

In this work, the transfer learning technique using Alexnet pretrained network has been applied. The advantage of this technique is the use of an existing neural network without the need to build it from scratch. This network is adapted to the image classification task and it can learn an efficient feature representation. Our work consists of fine-tuning this network to adapt it to our used database by replacing the final layers, we can then train the network so that it learns feature representation for our new task. As a result, we will have database images with the probability of belonging to each class. This information can be deployed



(a) Africa (b) Beach (c) Bus (d) Dinosaur (e) Elephant

(f) Flower (g) Food (h) Horse (i) Monument (j) Mountain

Fig. 1. Some samples representing the 10 semantic classes of Wang database.

later in image retrieval task: similarity between images is calculated based on the distance between class membership probabilities of query image and database images.

The selected methods to compare the results of our proposal are *Walia et al.* (noted as Walia) [7], *Elalami* [19], *ODBTC* [13], *DDBTC* [14] and *LMCTP* [26]. Walia method is based on a combination of color, texture, and shape features using different fusion techniques. The Elalami method introduces an effective matching strategy in order to measure similarity between images. Both of ODBTC and DDBTC consist in generating a feature vector derived respectively from low complexity-Dither Block Truncation and Dot-Diffused Block Truncation Coding. Finally, a merged color space is created in LMCTP, from which local mesh color features are extracted.

4.2 Hardware and software

The experiments have been established on a machine with Ubuntu 64 bits operating system, 2,9 GiB RAM, Intel Core 2 QUAD CPU with a frequency of 2,40 GHz and NVIDIA Graphic card. The used programming language is MATLAB R2018a.

4.3 Images database

In this paper, we use Wang(Corel 1k) database which is available in its website [3]. This database contains 1000 images grouped into 10 semantic classes: Africa, Beach, Bus, Dinosaur, Elephant, Flower, Food, Horse, Monument, Mountain. 70% of images are used to train the neural network, where 30% are left for the test phase. Figure 1 shows some samples of this database.

4.4 Results

We have employed several well known measures to measure the performance of each approach in order to establish a comparison between the different competi-

[3] http://wang.ist.psu.edu/docs/related/

tor methods from a quantitative point of view. The selected measures are the precision (P) and the recall (R). These measures provide a real number between 0 and 1, where higher is better, and they are given by the following equations:

$$P = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ number\ of\ retrieved\ images} \tag{3}$$

$$R = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ number\ of\ relevant\ images} \tag{4}$$

In order to check the goodness of our proposal, our approach has been compared to other CBIR classic approaches and the retrieval process has been established using two different comparisons. A first comparison, that we note as *Fixed number of retrieved images*, has been carried out. In this comparison the number of images to retrieve is fixed previously. The parameter k represents this number and we have considered $k = 20$ in this comparison. The most k similar images are selected based on their distances with the query image. So that, according to the lowest distances belong to the most similar images.

The performance measures of our proposal are reported in Table 1, where precision and recall values are shown for each class. As it can be observed the proposal achieves a high precision, although the recall is low.

Table 1. Precision and recall performances of our approach by using a k fixed number of retrieved images. In this case, $k = 20$.

Classes	Precision	Recall
Africa	0.9333	0.2667
Beach	0.9000	0.2571
Bus	1.0000	0.2857
Dinosaur	1.0000	0.2857
Elephant	1.0000	0.2857
Flower	0.9667	0.2767
Food	0.9683	0.2767
Horse	1.0000	0.2857
Monument	0.9667	0.2762
Mountain	0.8383	0.2395
Overall	0.95733	0.2735

On the other hand, Table 2 shows the comparison with other CBIR classic approaches. As it is shown, our convolutional neural network approach outperforms significantly these approaches in terms of precision.

Moreover, our experiments have shown that the number of images retrieved affects significantly the recall, which is not the case for precision. Figure 2 shows variation of precision and recall values in terms of the number of images retrieved. This number varies between 10 to 70 which is the maximum number of relevant images that can be retrieved (Number of instances in each class in

Table 2. Comparison between our approach and other traditional approaches in terms of precision. A k fixed number of retrieved images has been used. In this case, $k = 20$. The best result is highlighted in **bold**.

Approach	Precision
Walia [7]	0.783
Elalami [19]	0.761
ODBTC [13]	0.779
DDBTC [14]	0.792
LMCTP [26]	0.765
Our method	**0.957**

learning database). We can see that precision keeps almost the same value even with a higher number of retrieved images which reflects the effectiveness of our approach. However recall is increasing notably which means that more relevant images are being retrieved. Precision and recall are equal when the number of retrieved images is 70 which is the same as the number of relevant images.

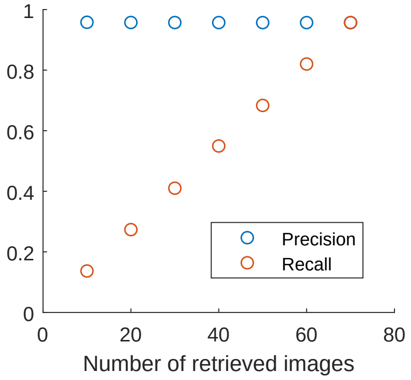


Fig. 2. Precision and recall performance of our proposal depending on the number of retrieved images parameter k.

Furthermore, the obtained results can be observed from a qualitative point of view. Figure 3 illustrates a user query with the relevant images returned by our system. In this figure, the image from the first row represents the query and the remaining images represent the retrieved images. For this query the obtained precision by our approach is perfect (so that, 1).

In addition, a second comparison, noted as *Defined threshold for the distance*, has been performed. In this case, for each query image, the number of retrieved images is determined based on a threshold τ. The distance between the query and all database images is calculated and then relevant images are selected. So

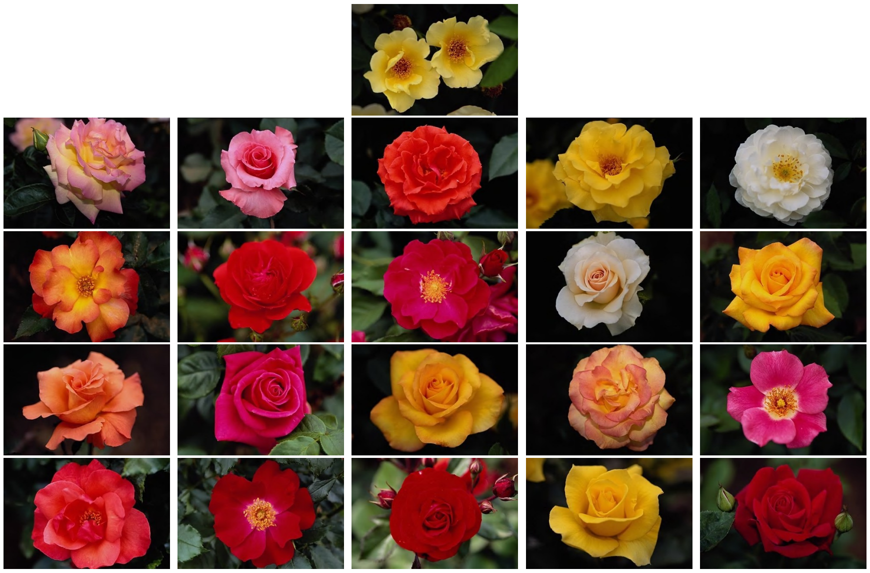


Fig. 3. An example of a query image from Flower class and retrieved images using our approach. The image from the first row represents the query and the remaining images represent the retrieved images.

that, the distance between a selected image and the query must be less than the defined threshold.

In this comparison, the definition of the threshold value is based on the precision, where the chosen threshold is the one that maximizes it. Figure 4 shows obtained precision values in terms of the used threshold.

After that, the precision and recall performances of our approach are reported in Table 3. Compared to the first comparison, it can be observed that the precision is practically the same but the recall has increased in a remarkable way. Thus, the used threshold allows our system to retrieve most of relevant images in the database.

5 Conclusion

A new content based image retrieval method by employing convolutional neural networks has been presented in this work. It uses a trained neural network in order to obtain the probabilities that an image represents an object class. Given a query image, two ways can be defined to provide the most similar images. In the first proposal, the output system is the k most similar images in the database, while in the second proposal the system supplies images which have a probability vector distance lower than a threshold τ.

The retrieval capabilities of the proposal have been tested in the experimental section. Well-known state-of-art methods, dataset and measures have been

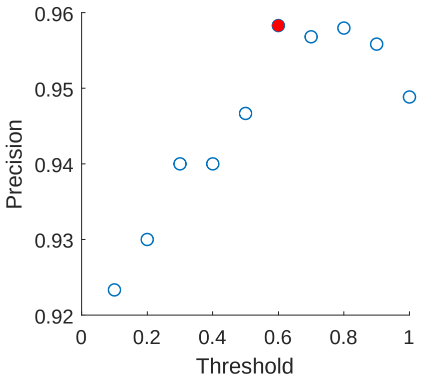


Fig. 4. Precision in terms of some experimented threshold values t. The red circle indicates the maximum reached precision which equals 0.9583 where the corresponding threshold value τ is $\tau = 0.6$.

Table 3. Precision and recall values using a defined threshold. Precision and recall performances of our approach by using a defined threshold τ for the distance. In this case, $\tau = 0.6$.

Classes	Precision	Recall
Africa	0.9333	0.9010
Beach	0.9000	0.9000
Bus	1.0000	1.0000
Dinosaur	1.0000	1.0000
Elephant	1.0000	1.0000
Flower	0.9667	0.9667
Food	0.9667	0.9667
Horse	1.0000	1.0000
Monument	0.9662	0.9667
Mountain	0.8500	0.8338
Overall	0.9583	0.9535

chosen to compare the performance. Quantitative results exhibit the goodness of the approach.

Acknowledgments

This work is partially supported by the Ministry of Economy and Competitiveness of Spain under grant TIN2014-53465-R, project name Video surveillance by active search of anomalous events. It is also partially supported by the Autonomous Government of Andalusia (Spain) under projects TIC-6213, project name Development of Self-Organizing Neural Networks for Information Technologies; and TIC-657, project name Self-organizing systems and robust estimators for video surveillance. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPU. Authors are also immensely grateful for ERASMUS+ program, CEI.MAR (Campus de Excelencia International del Mar), and University of 20 August 1955 for making this collaborative work possible.

References

1. Alec Thompson, Nathan George, Michael Gennert, Joseph Beck: Deep q-learning for humanoid walking (2016)
2. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (2012)
3. Allen Huang, Raymond Wu: Deep learning for music. arXiv preprint arXiv:1606.04930 (2016)
4. Andrej Karpathy, Li Fei-Fei: Deep visual-semantic alignments for generating image descriptions. The IEEE Conference on Computer Vision and Pattern Recognition (2015)
5. Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, Jason Yosinski: Plug & play generative networks: Conditional iterative generation of images in latent space. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich: Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (2015)
7. Ekta Walia, Aman Pal: Fusion framework for effective color image retrieval. J. Vis. Commun. Image R (6) (2014)
8. Florian Colombo, Samuel P. Muscinelli, Alexander Seeholzer, Johanni Brea, Wulfram Gerstner: Algorithmic composition of melodies with deep recurrent neural networks. Proceeding of the 1st Conference on Computer Simulation of Musical Creativity (2016)

9. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, Brian Kingsbury: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine (6) (2012)
10. Guillaume Lample, Devendra Singh Chaplot: Playing fps games with deep reinforcement learning. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (2017)
11. Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, Vinay Shet: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2014)
12. Ji Wan, Dayong Wang3, Steven C.H. Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, Jintao Li: Deep learning for content-based image retrieval: A comprehensive study. Proceedings of the ACM International Conference on Multimedia (2014)
13. Jing-Ming Guo, Heri Prasetyo: Content-based image retrieval using features extracted from halftoning-based block truncation coding. IEEE Transactions on Image Processing (3) (2015)
14. Jing-Ming Guo, Heri Prasetyo, Nai-Jian Wang: Effective image retrieval system using dot-diffused block truncation coding features. IEEE Transaction on Multimedia (9) (2015)
15. K. Kranthi Kumar, T. Venu Gopal: A novel approach to self order feature reweighting in cbir to reduce semantic gap using relevance feedback. International Conference on Circuits, Power and Computing Technologies (2014)
16. Karen Simonyan, Andrew Zisserman: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015)
17. Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, Chu-Song Chen: Deep learning of binary hash codes for fast image retrieval. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2015)
18. Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, Karol Zieba: End to end learning for self-driving cars. Computer Vision and Pattern Recognition (2016)
19. M.E. ElAlami: A new matching strategy for content based image retrieval system. Applied Soft Computing (2014)
20. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros: Image-to-image translation with conditional adversarial networks. Computer Vision and Pattern Recognition Proceeding (2017)
21. Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi: Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv:1702.07825 (2017)
22. Shusen Zhou, Qingcai Chen, Xiaolong Wang: Active deep networks for semi-supervised sentiment classification. Neurocomputing (2013)
23. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
24. Xiang-Yang Wang, Bei-Bei Zhang, Hong-Ying Yang: Content-based image retrieval by integrating color and texture features. Multimed Tools Appl (3) (2014)
25. Yaroslav Ganin, authorDaniil Kononenko, Diana Sungatullina, Victor Lempitsky: Deepwarp: Photorealistic image resynthesis for gaze manipulation. 14th Proceedings of European Conference on Computer Vision (2016)

26. Yesubai Rubavathi Charles, Ravi Ramraj: A novel local mesh color texture pattern for image retrieval system. International Journal of Electronics and Communications (3) (2016)
27. Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma: A survey of content-based image retrieval with high-level semantics. Pattern Recognition (1) (2007)
28. Yoshua Bengio, Aaron Courville, Pascal Vincent: Unsupervised feature learning and deep learning: A review and new perspectives. CoRR, abs/1206.5538 (2012)
29. Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh: Realtime multi-person 2d pose estimation using part affinity fields. Computer Vision and Pattern Recognition Proceeding (2017)