Image Compression Using Neural Networks: A Review

https://doi.org/10.3991/ijoe.v17i14.26059

Haval T. Sadeeq[1](✉), Thamer H. Hameed[2], Abdo S. Abdi[1], Ayman N. Abdulfatah[1]
[1]Duhok Polytechnic University, Dahuk, Kurdistan Region of Iraq
[2] University of Duhok, Dahuk, Kurdistan Region of Iraq
haval.tariq@dpu.edu.krd

Abstract—Computer images consist of huge data and thus require more memory space. The compressed image requires less memory space and less transmission time. Imaging and video coding technology in recent years has evolved steadily. However, the image data growth rate is far above the compression ratio growth, considering image and video acquisition systems' popularization. It is generally accepted, in particular, that further improvement in coding efficiency within the conventional hybrid coding system is increasingly challenging. A new and exciting image compression solution is also offered by the deep convolution neural network (CNN), which in recent years has resumed the neural network and achieved significant success both in artificial intelligent fields and in signal processing. In this paper we include a systematic, detailed and current analysis of image compression techniques based on the neural network. Images are applied to the evolution and growth of compression methods based on neural networks. In particular, the end-to-end frames based on neural networks are reviewed, revealing fascinating explorations of frameworks/standards for next-generation image coding. The most important studies are highlighted and future trends even envisaged in relation to image coding topics using neural networks.

Keywords—image compression, neural networks, artificial neural network, data compression

1 Introduction

Compression only minimizes the needed number of bits in order symbolize a video file or an image without dramatically affecting (original) input quality. One of the relevant compression areas is compression of images. The main aim of compression of the image is to reduce the amount of data required to symbolize an image. Even a standard digital image in a single band typically needs a large number of bits to display or store. The compression of images reduces the redundancies and irrelevances in the data. In contrast with the original image there are fewer bits needed to symbolize a compressed image. There are fewer bits needed to symbolize a compressed image in comparison to the original picture. The space used to store and bandwidth to convey the image is greatly reduced. To save the transmission time and bandwidth, we can save more images and transmit them in less time. Either image compression technique involves three main phases Mapper, Quantizer, and Encoder [1]. In the first step, Mapper converts the image to the frequency domain with special domain information. In order to map many to one theory Quantizer works. More values are quantified to fewer. The number of bits required to symbolize the image is therefore reduced. Finally, the result is coded during the encoding process to further improve the compression. To re-create the original image, several procedures are mapped namely; decoding, dequantization, and inverse mapping. This is called decompression [2], [3]. Compression and decompression block diagram are shown in Figure 1.

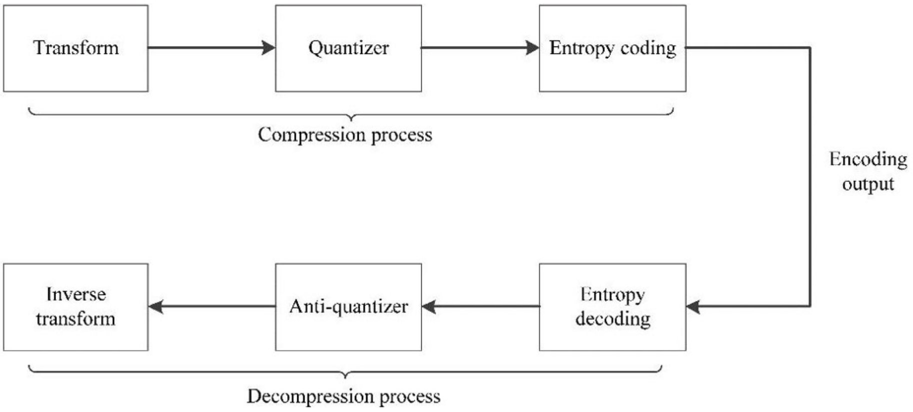

Fig. 1. Compression and decompression block diagram

In digital images, three fundamental redundancies are common:

1.1 Coding redundancy

The same number of bits is used in the normal image for more likely symbols and for lower probability symbols. We may boost this by allocating less bits to gray levels that seem more likely and more bits to less likely grays. In redundancy, there is no correlation between pixels [4].

1.2 Psycho-visual redundancy

There is material in digital images which is not relative to ordinary visual processing. Psychovisually redundancy is the name of this unimportant knowledge. The reduction of psychovisual redundancy leads to a lack of quantitative knowledge. The content of the image cannot be detected by the human eye [5].

1.3 Interpixel redundancy

Unrecognizing and utilizing data is known as redundancy of interpixel. If a value of its neighboring (subsequent) pixels can be predicted, the table would be redundant with interpixel values. The resolution of the picture influences the interpixel redundancies.

Redundancy between pixels of statistical dependence, particularly between adjacent pixels [6].

2 Image compression techniques

The compression techniques of the image [7] are usually known as a Lossy and a Lossless. Table 1 below presents different strategies for each group. The data would be lost if there is a loss compression, as its name implies. The image restored isn't the same as the image of the input. Figure 2 portrays the Lossy compression block drawing. Although the restored image is a complete replication of the original image with lossless techniques. No data loss occurs; decompressed image is identical to uncompressed one. Lossless methods of compression typically minimize file size for complex images by about ten percent. Health photographs or pictures in the courts are used. Lossless compression methods for simple images can provide significant compression. The block diagram is seen in the Figure 3.

Table 1. Classification of compression techniques

Lossy Techniques	Loseless Techniques
Transform coding	Run-length encoding
Vector quantization	Entropy coding
Fractal coding	Preditiction coding
Block transform coding	Huffman coding

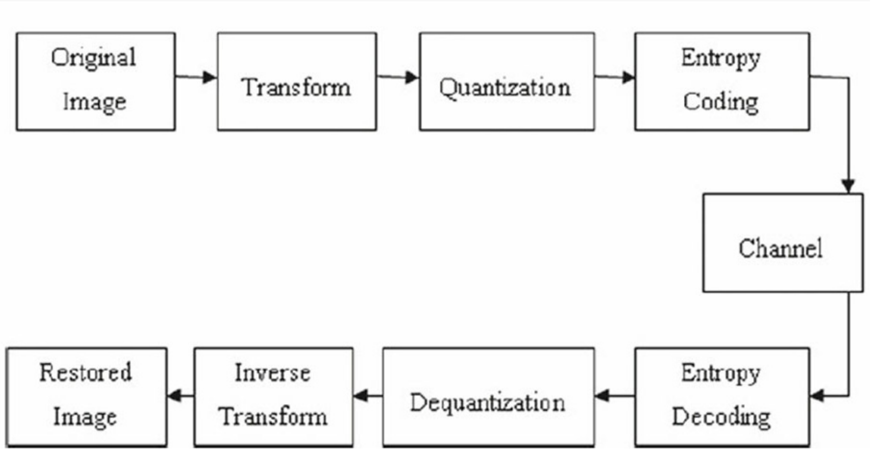

Fig. 2. Lossy image compression

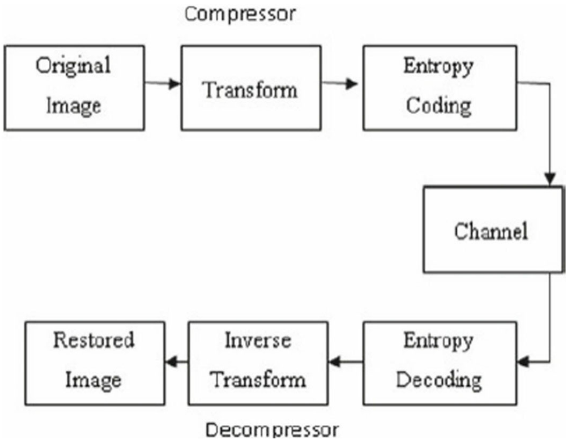

Fig. 3. Lossless image compression

2.1 Lossy techniques

Transform encoding. It is a sort of compression of relevant data like image or audio signals. It translates pixel image values from spatial domain to frequency domain. More information is given on the image with a few coefficients and most are low or zero. In order to map the pixel value of the image to a series of coefficients, a reversible linear transformation, such as Discrete Fourier or Discrete Cosine Transform is used. The coefficients are quantified and coded further. As much knowledge as possible will be gathered in a high-quality transformation in a limited number of coefficients. Then the quantization technique is used to extract coefficients that contain the least quantity of information. In this method, an input image of size N to N is first divided into a number of n to n that is not overlapping, then transformed into arrays (N/n)2, each sizes array n to n. Each block is separately transformed [8].

Vector quantization. Quantization operates on map many to one base. It maps a value group to one. In addition, the methods of quantization are divided into two types: scalar and vector quantization. scalar conducts a single mapping set for each value. Vector quantization is the multidimensional creation of scalar quantization. In both spectral and spatial domains, VQ can be implemented. The theory of data tells us that vector quantization can achieve better compression than scalar quantization. This technique constructs a dictionary of vectors known as vectors of code. The input image is divided into unrelated, image-vector-marked blocks. The index of the closest matching vector in the dictionary encodes any image vector [9].

Fractal coding. It assumes that a portion of an image is generally identical to its neighboring parts of the input image. These sections are called "fractal codes"; they are translated into mathematical data. The encoded image is restored by using these fractal codes. By another way, metaheuristic algorithms such as [10] [11] [12] represent a set of approaches used to solve hard optimization tasks with lucid resources consumption. They are characterized by their fast convergence and reduction of research complexity. It is worth noting that [13] and [14] have used metaheuristics for fractal image compression, and the performed experiments showed their effectiveness in the resolution of such problems.

Block transform coding (BTC). In BTC, the image input is divided in blocks where the size 8 pixels to 8 are used in every row. The coding of the block transform benefits from the correlation of the block pixels. Each block is modified by virtue of this. Finally, every block is individually quantified and coded [15].

2.2 Lossless techniques

Run-length encoding (RLE). Data is replaced in the RLE with a duo of (length, value), where the number of repetitions is called 'length' and the repeated value is referred to as 'value' This method is particularly used to compress large-scale images because in common, gray-size images, long-term value is not common. It is intended to broken down the gray image into the bit planes, and individual bit planes are individually compressed. One of the variants of run time coding is an efficient run-length coding mechanism [16].

Entropy encoding. Entropy is the lowest number of bits required to represent a symbol in the compression frame (the total length of the code for all symbols). The explicit features of the medium are not based. Every individual symbol in the input is assigned a unique prefix-free code. A prefix-free codeword for the variable length is used to replace each fixed-length input symbol. The codeword length is almost proportional to the probability logarithm. The most widely used symbols therefore take shorter codes [17].

Predictive coding. This technique is based primarily on the difference between the initial and expected values. It's often referred to as Differential Pulse Code Modulation (DPCM). It is not necessary to decompose the input image into a group of bit planes. Only new information is coded in each pixel after extracting the interpixel similarities. The variations between the original and expected value of the pixel are measured as new data. The system consists of a matching predictor in two phases: an encoder and a decoder. The predictor will generate the likely pixel value depending on the number of preceding inputs. The predicted value is coded to produce the next element of the data stream with a variable code length [18].

Huffman coding. In 1952, D. Huffman created the code of Huffman. It is a code of minimal duration. This means that the Huffman algorithm generates code as close as possible to the lower bound, entropy. The technique results in an unprecedented (or variable) code in which terms can vary in size [19] [20].

2.3 Assessment of compressed images

The compression ratio is the compression calculation of the image. Different factors including mean squared error and peak noise signal are determined by the compression efficiency.

There are also several other techniques; PSNR and MSE are mostly used because they are simple to measure [21] [22].

Mean square error (MSE). This is calculated by the mean difference between compressed and input (original) pixel square intensities. The MSE is indicated by the following equation.

$$\text{MSE} = \frac{1}{m*n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[g'(p,q)-g(p,q)]^2 \tag{1}$$

In these instances, the pixels of *g'*(*p, q*) and *g*(*p, q*) respectively are the values of the restored images. And *m & n* are the spatial domain number of rows and columns [23].

Peak signal to noise ratio (PSNR). PSNR is the typical way in which fidelity measurement is carried out [24] [25]. The term PSNR is the relation between the highest signal value of deforming noise, which changes its representation value and the signal value of the deforming noise. In terms of logarithmic decibel scale the PSNR is usually represented. In decibels, PSNR is computed (dB). In general, a large difference is said to be 0.5–1 dB. The PSNR's mathematical representation is:

$$\text{PSNR} = 10\log_{10}\left(\frac{m*n}{\text{MSE}}\right) \tag{2}$$

where MSE is Mean Square Error, and *m & n* are image rows & columns in spatial form.

Compression ratio (CR). Is a further (extra) measurement metric for compression measurement. It is the relative sum between the bits needed by the image input to the bits required by the compressed image. The following, is the compression ratio equation [26].

$$\text{C}_\text{R} = \frac{\text{Uncompressed (original) image size}}{\text{Compressed image size}} \tag{3}$$

2.4 Image formats

Images are divided into various categories such that certain image formats support lossless compression and lossless support. Table 2 tabulates different image formats and supporting techniques.

Table 2. Different image formats

Lossy Compression	Loseless Compression
JPG	GIF
JPEG	PNG
JPEG 2000	TIF
	BMP
	RAW

JPG. JPG analyzes the image and eliminates detail that cannot be seen by a human eye. It is used in continuous tone photographs and pictures. The data is saved in 24-bit color. The amount of compression can be changed [27].

JPEG. Is a perfect way to store 24-bit photographic images, abbreviated as a Joint Photographic Expert Group. It is used extensively in multimedia and online applications. JPEG compresses the image in such a way that the detail is lost. Also for the encoding of video is JPEG. Not suitable for diagrams and graphics [28] [29].

JPEG 2000. In JPEG the compression of the image is based on DCT coefficients, while in JPEG 2000 wavelet method is used. It was formed by the "*Joint Photographic Experts Group committee*" in the year 2000. JPEG 2000 addresses loss and lossless strategies [30] [31].

GIF. Graphics Interchange Format is the GIF's full format. The 8-bit colors are allowed. It is also used for (black and white) text and images in gray scale. The biggest downside of GIF's is that images with over 256 colors cannot be used, but most color pictures have over 256 colors, i.e. 24 bits per pixel. In animated pictures it is commonly used. [32] [33].

PNG. PNG is Portable Network Graphics' short form. This format is used for image compression without loss. The PNG file format replaces the GIF file format as compression is 10–30% higher than the GIF. PNG creates smaller files which allows for more colors. PNG supports partial transparency. There are two variants of PNG-24 (224 = 16777216 supports) and PNG-8 (28 = 256 colors supported) [34] [35].

TIFF. It stands for Tagged Image File Format short form. It can be considered as lossless format. The format is used primarily in photography and desktop publishing, because of its extremely high quality. Saving a total of 8, 16 bits for a total 24 and 48 bits respectively per (red, green, blue). Compression of TIFF in web applications is relatively poor and not used. They are particularly used in massive sizes in high-quality prints [36].

BMP. The BMP is a short for the Bitmap Image file. The raster graphic data is included with the BMP file. Showing devices do not impact this knowledge. This is not important to see the BMP file image for a graphics adapter. It is typically used with compression techniques without loss. It is also referred to as System Independent Bitmap (DIB short) or a bitmap. It primarily stores digital images on the OS/2 and Microsoft Windows operating systems in a wide variety. It can store digital images both monochrome and 2-D color [37] [38].

From above, the key techniques for compressing images are transformed and predicted among the different coding frameworks. JPEG is the common image compression standard consisting of the fundamental modules of the transform/prediction as shown in the Figure 4. The input image in JPEG is divided into 8_8 blocks that are not overlapped, all of which will be translated into block DCT frequency (BDCT). A binary stream is then compressed with DCT coefficients by quantizing and entropy coding for each transformed block [39] [40].

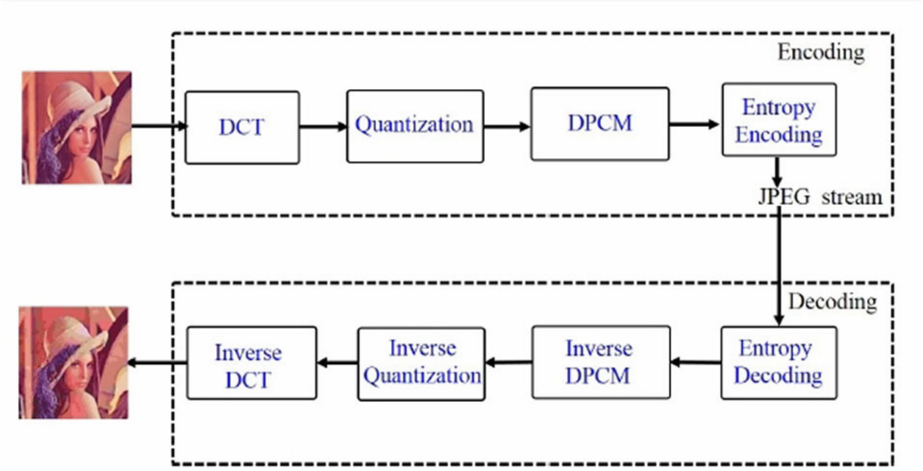

Fig. 4. JPEG compression

3 Neural networks

Also known as artificial neural networks (ANN) are one of the most common regression and classification models for machine learning and application in computer literature. In the sense of non-linear transformation and classification, ANN has shown good skills. The network comprises multiple layers of basic units called neurons that communicate through weighted connections with each other. The neurons are triggered by means of weighted neuron connections. In order to achieve non-linearity, all intermediate layers are always activated [41] [42]. In tasks like image recognition neural networks have achieved great success; handwriting recognition and optical character [43] [44]. In fact, the recent explosive advancement of deep learning [45] has resulted in the neural networks that are composed of several different layers of learners. In a task of image recognition, for example, where the machine has to classify what types of objects are in an image one layer might learn where the lines are in an image, while another layer could learn how these lines organize to represent different forms (e.g., books vs. people vs. pets).

Neural networks are a type of supervised learning based on biological structures and mechanisms of the human brain. The neural networks generate predictions in the layers of a set of connected nodes and neurons as stated above. As the neurons of this layer only accept variables of the data set as an input, the first layer is called the input layer. The last layer is the output layer since the final prediction is output (s). Hidden layers are between the input and the outputs, because the output in the network is only relevant. The neurons generate a weighted amount from their input and transform the weighted amount by means of some kind of nonlinear function such as the logit, hyperbolic tangent or the linear corrected function. In the next layer of the network, the measured value from the function is passed on the weighted sum. Knowledge flows from the input layer through the cached layers into the output layer through the neural network in one direction. If the output layer has been reached, it is stored and translated to predictions [46].

Neural networks may have multiple hidden layers, and the functions used inside the neurons can differ depending on the complexity of the data for which predictions are needed. In principle, a cached neural layer and an output layer with enough neurons are enough to learn about any binary task, and two layers plus an output layer will get close to each return task [47]. Tens of hidden layers can be used to support the discovery of deep learning patterns within the data in the becoming more common field of deep learning. The overall neural network architecture defines which neurons in subsequent layers feed their output into other neurons. The predicted variable shape also controls the number of neurons in the output layer. The final output layer consists particularly of one neuron for regression tasks (continuous results) and binary classification tasks (dichotomic result). Alternatively, the final output layer consists of a neuron per potential value for multinomial classification tasks, where there are more than two values in the categorical outcome variable. In this case, the expected class is the neuron with the highest value. The basic neural network architecture is illustrated in Figure 5. ANN is consisting of one input layer, one output layer and different hidden layers, each containing various neurons.

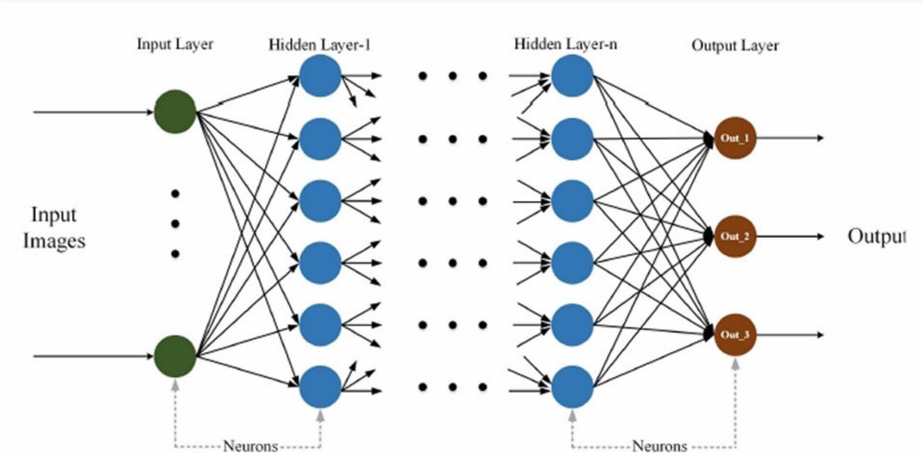

Fig. 5. Neural network architecture

4 Image compression using neural network

In this section we present machine learning methods for image compression, mainly originating in the late 1980s, especially from a neural network point of view. This section consists of the history of neural network techniques, mainly the Multilayer Perceptron Network (MLP), the Random Neural Network, the Recurrent Networks (RNN) and the Convolutional Neural Network (CNN). The last section will present the recent development of the techniques for image coding with generative adversarial networks (GAN) [48].

Neural network is a recent compression tool because data processing is in parallel manner and thus requires less time, and it is general performance is superior to any other technique. It is therefore important to transform a neural network image information efficiently. The data can be transformed by techniques like PCA based on factorizing techniques developed in linear algebra [49].

4.1 Multi-layer perceptron based image coding

MLP is made of a neuronal input (nodes) layer, a variety of hidden neuron layers, and an ultimate neuron output layer [50]. The output of any neuron I in the MLP is referred to as,

$$h_i = \sigma\left(\sum_{j=1}^{N} w_{ij}\, x_j + c_i\right), \tag{4}$$

where σ is the activation function, c_i denotes the linear transform biasterm and wij indicates the weight adjustable parameter which is the layer connection. The theoretical study has shown that the MLP developed with one hidden layer can bring about arbitrary accuracy of any continuous computable function [51]. This property shows scenarios such as reduction of sizes and compression of data. The initiative of using MLP to establish unitary transformations for spatial data as a whole is the compression of image. [52] proposed a complete image compression method based on nonlinear transform coding, and a framework to optimize it end-to-end for rate–distortion performance. The compression method offers improvements in rate–distortion performance over JPEG and JPEG 2000 for most images and bit rates. More remarkably, although the method was optimized using mean squared error as a distortion metric, the compressed images are much more natural in appearance than those compressed with JPEG or JPEG 2000, both of which suffer from the severe artifacts commonly seen in linear transform coding methods. Consistent with this, perceptual quality exhibits substantial improvement across all test images and bit rates. The authors believe this visual improvement arises because the cascade of biologically-inspired nonlinear transformations in the model have been optimized to capture the features and attributes of images that are represented in the statistics of the data.

In order to find an optimized combination of binary code, a decomposition/decision neural network, is then utilized by authors to solve the optimization problem. Every 8_8 patch of an image was compressed with retrospect propagation by a fully linked neural network with 16 hidden units in [53]. This technique, however, specified the neural network parameters for the specific number of binary codes, that can hardly be adapted in optimal form for variable compression.

[54] proposed a novel iris recognition method-based Multilayer perception neural network (MLPNN) and particle swarm optimization algorithm (PSO) to classify iris images. The 2D Gabor kernel algorithm was used for feature extraction. The results of testing on the CASIA-iris V3 database and UCI machine learning repository databases indicate that the hybrid MLPNN–PSO algorithm is an effective, appropriate, stable, robust, and competitive recognition method for human iris recognition.

[55] focused on the study of dimension reduction. After using a multilayer neural network to extract image features, the Principal Component Analysis (PCA) algorithm is used to achieve dimension reduction. Specifically, they first leverage a multilayer neural network to extract image features. Then they introduced and leveraged the PCA algorithm to achieve dimension reduction. Aiming at the problem that it is difficult to process high-dimensional sparse big data based on the PCA algorithm.

4.2 Random neural network based image coding

In 1989, a new class was developed of the random neural network. Random neural networks operate in a different way than the MLP methods described above, where signals are optimized by the backpropagation process in the field of spaces. Random neural network signals are transmitted as unit amplitude spikes. These neurons communicate as a Poisson mechanism. The behavior of random neural networks have been studied by some theoretical results [56].

[57] proposed a system that can be used to identify the image of Pekalongan batik using Backpropagation Artificial Neural Network. Sogan's motive is the result with the highest level of accuracy that reaches 91.2 %% in testing the value of age = 100 and the level of learning = 0.03. the second level of accuracy was 89.6% with age 100 and learning level 0.02 in Jlamprang batik. The last level of 87.2% was obtained from the Cap Combination batik and Tiga Negeri batik with the age of 100, the learning level was 0.01 and 0.04. High and low accuracy values can be caused by structured motives and the sharpness of batik colors will help increase the value of accuracy.

In [58] the implementation of back propagation neural network algorithm on image compression system with good performance has been demonstrated. The back propagation neural network has been trained and tested for the analysis of different images. It has been observed that the convergence time for the training of back propagation neural network is very faster. Different attributes of compression such as compression ratio, peak signal to noise ratio, bits per pixel are calculated. It has been observed that there is significance change in compression ratio from .99 to .9556 in case of Cameraman image. It has also been observed that there is significance improvement in peak signal to noise ratio from 19.3181 to 20.722 in case Cameraman.

4.3 Convolutional neural network based coding

CNN's latest success in high level computer vision tasks, including classifying images, is outstrips conventional algorithms with a wide margin [59]. It also achieves remarkable results, for example super-resolution and compression artifact reductions, also in many low-level computer vision tasks. In order to characterize the link between neighboring pixels, the CNN undertakes a convolutional operation with the cascade convolution operations well matched with the hierarchical statistical properties. Moreover, the local receptive areas as well as the common weight of the convergence operations often lower the trainable CNN parameters, reducing the chance of overfitting significantly. Inspired by powerful CNN pictorial representation, several works were done to explore the feasibility of compression of the CNN-based loss image. But implementing the CNN model in the end-to-end compression is difficult straightforward. In general, CNN training relies on the back-propagation and stochastic gradient descent algorithm, which require that the loss function is almost uniformly different from qualified parameters, such as convergence weights and bias. Thanks to the image compression quantization module, it generates virtually everywhere zero gradients, avoiding CNN updating parameters. Moreover, it is difficult to optimize the classical rate-distortion for a CNN-based compression frame. Since end-to-end CNN formation requires an adjustable loss feature, but the rate must be estimated based on the population distribution of whole quantifiable bins. [60] first implemented a scalar quantization assumption, with an end-to-end, optimized CNN image compression system. By using a pyramidal function fusion structure in an encoder and a CNN-based decoder post processing filter.

[61] have presented a 12-layer deep convolutional neural network for compression artifact suppression in JPEG images with hierarchical skip connections and trained with a multi-scale loss function. The result is a new state-of-the-art ConvNet achieving a boost of up to 1.79 dB in PSNR over ordinary JPEG and showing an improvement of up to 0.36 dB over the best previous ConvNet result. We have shown that a network trained for a specific quality factor is resilient to the QF used compress the input image a single network trained for QF 60 provides a PSNR gain of more than 1.5 dB over the wide QF range from 40 to 76. The obtained results are also qualitatively superior to those of existing ConvNets.

[62] proposed a Deep Dual-Domain (D3) based fast restoration model to remove artifacts of JPEG compressed images. It leverages the large learning capacity of deep networks, as well as the problem-specific expertise that was hardly incorporated in the past design of deep architectures. Accordingly, the authors take into consideration both the prior knowledge of the JPEG compression scheme, and the successful practice of the sparsity-based dual-domain approach. The successful combination of both JPEG prior knowledge and sparse coding expertise has made D3 highly effective and efficient.

4.4 Recurrent neural network based coding

Contrary to the above-noted CNN architecture, RNN is a neural memory class to store current comportments. Memory units in RNN, in particular, have relations with themselves, which have previously transmitted transformed data from output. RNN adjusts the actions of the present forward phase in order to respond to the context of the current input by utilizing this stored information.

[63] first suggested a scaled-additive coding framework to reduce the number of coding bits rather than approx. the estimate rate in CNN by using an RNN-based image compression scheme.

A set of full-resolution lossy image compression methods based on neural networks was proposed by [64]. Each of the architectures can provide variable compression rates during deployment without requiring retraining of the network: each network need only be trained once. All of the architectures consist of a recurrent neural network (RNN)-based encoder and decoder, a binarizer, and a neural network for entropy coding.

[65] proposed a method for lossy image compression based on recurrent, convolutional neural networks that outperforms BPG (4:2:0), WebP, JPEG2000, and JPEG as measured by MS-SSIM. They introduced three techniques: hidden-state priming, spatially adaptive bit rates, and perceptually-weighted training loss and showed that they boost the performance of our baseline recurrent image compression architecture.

Recent article in [66] focused on developing deeper and more complex networks, which significantly increased network complexity. Two effective blocks are developed: analysis and synthesis block that employs the convolution layer and Generalized Divisive Normalization (GDN) in the variable rate encoder and decoder side. The proposed network utilizes a pixel RNN approach for quantization. Furthermore, to improve the whole network, they encode a residual image using LSTM cells to reduce unnecessary information. Experimental results demonstrated that the proposed variable-rate framework with novel blocks outperforms existing methods and standard image codecs.

4.5 Generative adversarial network based coding

GANs for short, GANs can be defined as the unsupervised learning task in the field of machine learning that automatically requires the discovery and learning of regularities and models in the input data, enabling the creation or production of new examples that may have been extracted from the original data set. GAN's are an intelligent way to construct a generative model by posing the issue with two substructures as a supervised learning problem: the model generator used to produce more features, and the model that tries to classify features as true or false examples (generated).

[67] are proposing one of interesting works and highly optimized GAN-based image compression, with a remarkable improvement of the compression ratio and the use of large parallel computing cores from GPU in real time, networks compress the input image into a very compact function space and the generative network is used to reconstituting the decoded image from the features. The clearest distinction between GAN and RNN or CNN-based image compression is the inclusion of an opposite loss that dramatically improves the subjective quality of the reconstructed image. The generative network and opposing network are jointly trained to increase the efficiency of the generative model significantly.

The GAN approach in [68] greatly enhances compression, for example producing standardized images for all compressed file quality levels, 2.5 times smaller than JPEG and JPEG 2000, 2x smaller than WebP and 1.7x smaller than BPG. In this sense, quality can be calculated using MS-SSIM, while PSNR metrics still don't work. The compression of the light field image (LF) could achieve substantial coding gains based on the progress made in GAN-based view synthesis by generating the missing views with sampled background views in LF. GAN creates content, in fact, more compatible than basic textures with the semantics of the original content. In particular, we can see the differences in contents in specific textures while expanding reconstructed images.

[69] proposed a unified binary generative adversarial network (BGAN+) to simultaneously convert images to binary codes for both image compression and retrieval in a multi-task fashion and an unsupervised way. By restricting the input noise variable of generative adversarial networks (GAN) to be binary and conditioned on the features

of each input image, BGAN+ can simultaneously learn two binary representations per image: one for image retrieval and one for image compression. To equip the binary representation with the ability of accurate image retrieval and compression, we design a novel loss function. The results show that the proposed method outperforms the existing retrieval methods with significant margins and the multi-task strategy is beneficial for both tasks.

[70] tried to compact conceptuality by producing the most possible image semantic zed information. A GAN-based image compression system, targeting bitrates below 0.1 bpp, is studied in depth, allowing various content generation degrees. Currently, GAN compression is effective in narrow-domain images such as faces and still needs more research on natural-image modeling in general [71] [72] [73].

5 Discussion and recommendations

On the basis of the analysis, we believe that the neural network advantages are three-fold in picture and video compression. Firstly, the outstanding quality data adaptability of a neural network exceeds the signal processing model because the network parameters are based on a great deal of realistic details, while the models are designed by hand using previous imaging and visual expertise in the state-of-the-art coding standards. Second, the wider region of reception is commonly used in neural network models, which not only utilizes the information given in the vicinity, but also improves coding efficiency by utilizing samples from afar, but only use the conventional coding methods for neighboring and distant samples are difficult to use. Third, both the structure and the function that enables the joint compression optimization both for human view and machine vision analysis can be described by the neural network. However, only high compression efficiency for human viewing is achieved by the new coding standards. We intend to further research the deep learning image and video compression in the representation and distribution of image and videos with better quality and less bit rates and the memory and computational efficiency in the realistic photo. Computing and memory burdens are the greatest obstacle to the deployment of a deep learning picture. For higher performance, larger neural networks are typically taken into consideration with more layers and nodes but different network parameters' efficiency is not well explored.

6 Conclusion

Image compression is intended to pursue more efficient visual signals representation during retaining high-quality and growing significance in large-scale visual data. The neural network compression techniques, in particular the recent deep learning and video compression techniques, were discussed in this paper. The results are now available. The survey previously presented here shows that the new end-to-end image compression based on neural network is still in its infancy and is just overperforming the JPEG2000 and battling against HEVC.

On the basis of the discussion in this article, the neural network showed encouraging results for different compression tasks in the images and videos. Even if it is machine sophistication and memories usage are still troublesome, the neural network has achieved significant coding gains in addition to the cutting-edge video encoding frameworks due to its high efficiency of prediction or compact presentation of image and video signals.

7 References

[1] J. Nowaková, M. Prílepok, and V. Snášel, "Medical Image Retrieval Using Vector Quantization and Fuzzy S-tree," *J. Med. Syst.*, vol. 41, no. 2, 2017, doi: https://doi.org/10.1007/s10916-016-0659-2

[2] R. K. Netalkar, H. Barman, R. Subba, K. V. Preetam, and U. S. N. Raju, "Distributed Compression and Decompression for Big Image Data: LZW and Huffman Coding," *J. Electron. Imaging*, vol. 30, no. 5, 2021, doi: https://doi.org/10.1117/1.JEI.30.5.053015

[3] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-Art in 360 Video/Image Processing: Perception, Assessment and Compression," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 1, pp. 5–26, 2020, doi: https://doi.org/10.1109/JSTSP.2020.2966864

[4] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, 2021, doi: https://doi.org/10.1007/s11263-020-01419-7

[5] M. Krishna, G. Srinivas, and P. V. G. D. Prasad Reddy, "Image Smoothening and Morphological Operators Based JPEG Compression," *J. Theor. Appl. Inf. Technol.*, vol. 85, no. 3, pp. 252–259, 2016.

[6] Z. Chen, X. Hou, X. Qian, and C. Gong, "Efficient and Robust Image Coding and Transmission Based on Scrambled Block Compressive Sensing," *IEEE Trans. Multimed.*, vol. 20, no. 7, pp. 1610–1621, 2018, doi: https://doi.org/10.1109/TMM.2017.2774004

[7] A. J. Hussain, A. Al-Fayadh, and N. Radi, *Image compression techniques: A survey in lossless and lossy algorithms*, vol. 300. 2018. https://doi.org/10.1016/j.neucom.2018.02.094

[8] H. Zhang and L. Hu, "A Data Hiding Scheme Based on Multidirectional Line Encoding and Integer Wavelet Transform," *Signal Process. Image Commun.*, vol. 78, no. January, pp. 331–344, 2019, doi: https://doi.org/10.1016/j.image.2019.07.019

[9] P. P. Chavan, B. S. Rani, M. Murugan, and P. Chavan, "A Novel Image Compression Model by Adaptive Vector Quantization: Modified Rider Optimization Algorithm," *Sadhana – Acad. Proc. Eng. Sci.*, vol. 45, no. 1, 2020, doi: https://doi.org/10.1007/s12046-020-01436-9

[10] H. Sadeeq, A. Abdulazeez, N. Kako, and A. Abrahim, "A Novel Hybrid Bird Mating Optimizer With Differential Evolution for Engineering Design Optimization Problems," *Lect. Notes Data Eng. Commun. Technol.*, vol. 5, pp. 522–534, 2018, doi: https://doi.org/10.1007/978-3-319-59427-9_55

[11] H. T. Sadeeq, A. M. Abdulazeez, N. A. Kako, D. A. Zebari, and D. Q. Zeebaree, "A New Hybrid Method for Global Optimization Based on the Bird Mating Optimizer and the Differential Evolution," *Proc. 7th Int. Eng. Conf. "Research Innov. Amid Glob. Pandemic", IEC 2021*, no. August, pp. 54–60, 2021, doi: https://doi.org/10.1109/IEC52205.2021.9476147

[12] H. Sadeeq and A. M. Abdulazeez, "Hardware Implementation of Firefly Optimization Algorithm Using FPGAS," *ICOASE 2018 – Int. Conf. Adv. Sci. Eng.*, pp. 30–35, 2018, doi: https://doi.org/10.1109/ICOASE.2018.8548822

[13] R. Menassel, B. Nini, and T. Mekhaznia, "An Improved Fractal Image Compression Using Wolf Pack Algorithm," *J. Exp. Theor. Artif. Intell.*, vol. 30, no. 3, pp. 429–439, 2018, doi: https://doi.org/10.1080/0952813X.2017.1409281

[14] R. Menassel, I. Gaba, and K. Titi, "Introducing BAT Inspired Algorithm to Improve Fractal Image Compression," *Int. J. Comput. Appl.*, vol. 42, no. 7, pp. 697–704, 2020, doi: https://doi.org/10.1080/1206212X.2019.1638631

[15] S. Limuti, E. Polo, and S. Milani, "A Transform Coding Strategy for Voxelized Dynamic Point Clouds," *Proc. – Int. Conf. Image Process. ICIP*, pp. 2954–2958, 2018, doi: https://doi.org/10.1109/ICIP.2018.8451254

[16] H. H. Nuha, "Lossless Text Image Compression using Two Dimensional Run Length Encoding," *J. Online Inform.*, vol. 4, no. 2, p. 75, 2020, doi: https://doi.org/10.15575/join.v4i2.330

[17] U. Sharma, M. Sood, E. Puthooran, and Y. Kumar, "A Block-based Arithmetic Entropy Encoding Scheme for Medical Images," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 15, no. 3, pp. 65–81, 2020, doi: https://doi.org/10.4018/IJHISI.2020070104

[18] Z. Chen, S. Member, J. Xu, C. Lin, and W. Zhou, "Stereoscopic Omnidirectional Image Quality Assessment Based on Predictive Coding Theory," pp. 1–13.

[19] J. V. C. I. R, S. Yuan, and J. Hu, "Research on Image Compression Technology Based on Huffman Coding Q," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 33–38, 2019, doi: https://doi.org/10.1016/j.jvcir.2018.12.043

[20] M. Testolina, T. Ebrahimi, S. Diego, and U. States, "Evaluation," no. August, 2021, doi: https://doi.org/10.1117/12.2597813

[21] X. Min *et al.*, "Unified Blind Quality Assessment of Compressed Natural, Graphic, and Screen Content Images," *IEEE Trans. Image Process.,* vol. 26, no. 11, pp. 5462–5474, 2017. https://doi.org/10.1109/TIP.2017.2735192

[22] X. Zhang, W. Lin, and S. Wang, "Fine-Grained Quality Assessment for Compressed Images," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1163–1175, 2019, doi: https://doi.org/10.1109/TIP.2018.2874283

[23] D. Minnen, J. Ballé, and G. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression," no. Nips, 2018.

[24] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression Artifacts Reduction by a Deep Convolutional Network," pp. 576–584.

[25] K. Gregor and F. Besse, "Towards Conceptual Compression," no. Nips, 2016.

[26] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An End-to-End Compression Framework Based on Convolutional Neural Networks," pp. 1–13.

[27] G. K. Dziugaite and D. M. Roy, "A Study of the Effect of JPG Compression on Adversarial Images arXiv : 1608.00853v1 [cs.CV] 2 Aug 2016," no. Isba, 2016.

[28] N. Das *et al.*, "Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression," pp. 196–204, 2018. https://doi.org/10.1145/3219819.3219910

[29] X. Fu, X. Wang, A. Liu, J. Han, and Z. Zha, "Learning Dual Priors for JPEG Compression Artifacts Removal," pp. 4086–4095, 2019.

[30] L. Theis and W. Shi, "Lossy Image Compression with Compressive Autoencoders," pp. 1–19, 2017.

[31] F. Yang, L. Herranz, Y. Cheng, and M. G. Mozerov, "Slimmable Compressive Autoencoders for Practical Neural Image Compression," pp. 4998–5007.

[32] B. O. Ayinde, "A Fast and Efficient Near-Lossless Image Compression using Zipper Transformation arXiv : 1710.02907v2 [eess.IV] 10 Oct 2017," pp. 1–13.

[33] C. Oswald, E. Haritha, A. A. Raja, and B. Sivaselvan, "An Efficient and Novel Data Clustering and Run Length Encoding Approach to Image Compression," no. June 2020, pp. 1–16, 2021, doi: https://doi.org/10.1002/cpe.6185

[34] K. Nakanishi, S. Maeda, T. Miyato, and D. Okanohara, “Neural Multi-scale Image Compression,” 2018.
[35] M. Reconstruction, C. Framework, B. On, and P. N. G. Image, “Multiple Reconstruction Compression Framework,” doi: https://doi.org/10.5121/ijcsity.2019.7401
[36] A. J. Qasim, R. Din, F. Qasim, and A. Alyousuf, “Review on Techniques and File Formats of Image Compression,” *Bulletin of Electrical Engineering and Informatics,* vol. 9, no. 2, pp. 602–610, 2020, doi: https://doi.org/10.11591/eei.v9i2.2085
[37] C. Series, “Comparing Freeman Chain Code 4 Adjacency Algorithm and LZMA Algorithm in Binary Image Compression Comparing Freeman Chain Code 4 Adjacency Algorithm and LZMA Algorithm in Binary Image Compression,” 2021, doi: https://doi.org/10.1088/1742-6596/1783/1/012045
[38] S. Ezilarasi, “Enhanced AMBTC Based Adaptive Image Compression Technique,” vol. 7, no. 8, pp. 5690–5702, 2020.
[39] N. A. Kako, H. T. Sadeeq, and A. R. Abrahim, “New Symmetric Key Cipher Capable of Digraph to Single Letter Conversion Utilizing Binary System,” no. May, pp. 1028–1034, 2020, doi: https://doi.org/10.11591/ijeecs.v18.i2.pp1028-1034
[40] N. A. Kako, A. M. Abdulazeez, and H. T. Sadeeq, “Effect of Colored Noise on Neuron Membrane Size Using Stochastic Hodgkin-Huxley Equations,” 2021. https://doi.org/10.1109/IEC52205.2021.9476110
[41] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, “Gene Selection and Classification of Microarray Data Using Convolutional Neural Network,” *2018 Int. Conf. Adv. Sci. Eng.*, no. November, pp. 145–150, 2018. https://doi.org/10.1109/ICOASE.2018.8548836
[42] J. N. Saeed and A. M. Abdulazeez, “Facial Beauty Prediction and Analysis Based on Deep Convolutional Neural Network: A Review,” vol. 1, pp. 1–12, 2021. https://doi.org/10.30880/jscdm.2021.02.01.001
[43] M. R. Mahmood, “A New Hand Gesture Recognition System Using Artificial Neural,” no. April 2017, 2019.
[44] S. I. Saleem, A. M. Abdulazeez, and Z. Orman, “A New Segmentation Framework for Arabic Handwritten Text Using Machine Learning Techniques A New Segmentation Framework for Arabic Handwritten Text,” no. April, 2021, doi: https://doi.org/10.32604/cmc.2021.016447
[45] X. Zhang, S. Wang, Y. Zhang, and W. Lin, “High-Efficiency Image Coding via Near-Optimal Filtering,” vol. 24, no. 9, pp. 1403–1407, 2017. https://doi.org/10.1109/LSP.2017.2732680
[46] K. Andersson, M. Zhou, and G. Van Der Auwera, “HEVC Deblocking Filter,” vol. 22, no. 12, pp. 1746–1754, 2012, doi: https://doi.org/10.1109/TCSVT.2012.2223053
[47] C. Fu *et al.*, “Sample Adaptive Offset in the HEVC Standard,” vol. 22, no. 12, pp. 1755–1764, 2012. https://doi.org/10.1109/TCSVT.2012.2221529
[48] J. Lin, Y. Chen, Y. Huang, S. Lei, and T. He, “Motion Vector Coding in the HEVC Standard,” no. c, 2013. https://doi.org/10.1109/JSTSP.2013.2271975
[49] X. Zhang, R. Xiong, and W. Lin, “Low-Rank based Nonlocal Adaptive Loop Filter for High Efficiency Video Compression,” vol. 8215, no. c, pp. 1–12, 2016, doi: https://doi.org/10.1109/TCSVT.2016.2581618
[50] S. Ma, X. Zhang, J. Zhang, C. Jia, S. Wang, and W. Gao, “Nonlocal In-Loop Filter : The Future Way Towards Next-Generation Video Coding?” pp. 1–9, 2016. https://doi.org/10.1109/MMUL.2016.16
[51] C. Tsai *et al.*, “Adaptive Loop Filtering for Video Coding,” vol. 7, no. 6, pp. 934–945, 2013. https://doi.org/10.1109/JSTSP.2013.2271974
[52] J. Ball and E. P. Simoncelli, “E—o i c,” 2017.
[53] M. W. Gardner and S. R. Dorling, “Artificial Neural Networks (The Multilayer Perceptron)—A Review Of Applications In The Atmospheric Sciences,” vol. 32, no. 14, pp. 2627–2636, 1998. https://doi.org/10.1016/S1352-2310(97)00447-0

[54] N. Ahmadi and G. Akbarizadeh, "A Hybrid Robust Iris Recognition Approach Using Iris Image Preprocessing, 2D Hybrid Robust Iris Recognition Approach Using Iris Image Pre-processing, Two-dimensional Gabor Features and Multi-layer Perceptron Neural Network/PSO," no. May 2018, 2017, doi: https://doi.org/10.1049/iet-bmt.2017.0041

[55] J. V. C. I. R, J. Ma, and Y. Yuan, "Dimension Reduction of Image Deep Feature Using PCA Q," vol. 63, 2019, doi: https://doi.org/10.1016/j.jvcir.2019.102578

[56] K. Dimililer, "Backpropagation Neural Network Implementation for Medical Image Compression," *Journal of Applied Mathematics*, vol. 2013, 2013. https://doi.org/10.1155/2013/453098

[57] R. A. Surya, A. Fadlil, and A. Yudhana, "Identification of Pekalongan Batik Images Using Backpropagation Method Identification of Pekalongan Batik Images Using Backpropagation Method," 2019, doi: https://doi.org/10.1088/1742-6596/1373/1/012049

[58] S. S. Panda, M. S. R. S. Prasad, M. N. M. Prasad, and C. S. Naidu, "Image Compression Using Back Propagation Neural Network," pp. 74–78, 2012.

[59] S. A. Dianat, N. M. Nasrabadi, S. Venkataraman, and R. Inst, "A . m .," vol. 2, pp. 2793–2796, 1991.

[60] V. Quantization, O. Images, B. U. A. Neural-network, and C. Algorithm, "Vector Quantization Of Images Clustering Algorithm," no. October 1988, 2021, doi: https://doi.org/10.1117/12.968954

[61] L. Cavigelli, P. Hager, and L. Benini, "CAS-CNN: A Deep Convolutional Neural Network for Image Compression Artifact Suppression."

[62] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang, "D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images."

[63] E. Gelenbe, "Random Neural Networks with Negative and Positive Signals and Product Form Solution," *Neural Comput.*, vol. 1, no. 4, pp. 502–510, 1989, doi: https://doi.org/10.1162/neco.1989.1.4.502

[64] G. Toderici *et al.*, "Full Resolution Image Compression With Recurrent Neural Networks," *Proc. – 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017, pp. 5435–5443, 2017, doi: https://doi.org/10.1109/CVPR.2017.577

[65] N. Johnston *et al.*, "Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4385–4393, 2018, doi: https://doi.org/10.1109/CVPR.2018.00461

[66] K. Islam, L. M. Dang, S. Lee, and H. Moon, "Image Compression with Recurrent Neural Network and Generalized Divisive Normalization," pp. 1875–1879, 2021, doi: https://doi.org/10.1109/CVPRW53098.2021.00209

[67] F. Hai, K. F. Hussain, E. Gelenbe, and R. K. Guha, "Video Compression With Wavelets and Random Neural Network Approximations," *Appl. Artif. Neural Networks Image Process. VI*, vol. 4305, no. April, pp. 57–64, 2001, doi: https://doi.org/10.1117/12.420926

[68] Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: https://doi.org/10.1038/nature14539

[69] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, *Unified Binary Generative Adversarial Network for Image Retrieval and Compression*, vol. 128, no. 8–9. 2020. https://doi.org/10.1007/s11263-020-01305-2

[70] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational Image Compression With a Scale Hyperprior," *6th Int. Conf. Learn. Represent. ICLR 2018 – Conf. Track Proc.*, 2018.

[71] Y. He, "Application of Panoramic Image Technology in Distance Teaching System," *Int. J. Emerg. Technol. Learn.*, vol. 10, no. 6, pp. 27–31, 2015, doi: https://doi.org/10.3991/ijet.v10i6.4848

[72] L. Favario and E. Masla, "A New Architecture for Cross-Repository Creation and Sharing of Educational Resources," *Int. J. Emerg. Technol. Learn.*, vol. 12, no. 2, pp. 185–209, 2017, doi: https://doi.org/10.3991/ijet.v12i02.6058

[73] S. Zhu, "An Online Interaction Mode for International Trade Practice Course Under Network Environment," *Int. J. Emerg. Technol. Learn.*, vol. 12, no. 7, pp. 32–43, 2017, doi: https://doi.org/10.3991/ijet.v12i07.7219

8 Authors

Haval Tariq Sadeeq is a lecturer in the Information Technology Department at Duhok Polytechnic University in Duhok, Kurdistan Region of Iraq. Currently, he is a PhD student at the Technical College of Informatics/DPU. His interests are Soft Computing, Swarm Intelligence, Image Processing, Data Cryptography and Compression. E-mail: haval.tariq@dpu.edu.krd

Thamer Hassan Hameed is a lecturer at the University of Duhok, Duhok, in the Kurdistan Region of Iraq. His interests are Swarm Intelligence, Machine Learning and Image Processing. E-mail: thamer.hameed@uod.ac

Abdo Sulaiman Abdi is a lecturer in the Information Technology Department at Duhok Polytechnic University, Duhok, Kurdistan Region of Iraq. His interests are computer hardware, networking, and communications. E-mail: abdo.abdi@dpu.edu.krd

Ayman Nashwan Abdulfatah is a lecturer in the Information Technology Department at Duhok Polytechnic University in Duhok, Kurdistan Region of Iraq. his interests are DSP, Web programming, Computer Network. E-mail: ayman.nashwan@dpu.edu.krd

Article submitted 2021-08-06. Resubmitted 2021-10-12. Final acceptance 2021-10-12. Final version published as submitted by the authors.