Computers, Materials & Continua
DOI: 10.32604/cmc.2022.026602
Article

Tech Science Press

Low Complexity Encoder with Multilabel Classification and Image Captioning Model

Mahmoud Ragab[1,2,3,*] and Abdullah Addas[4]

[1]Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[2]Centre of Artificial Intelligence for Precision Medicines, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[3]Mathematics Department, Faculty of Science, Al-Azhar University, Naser City, 11884, Cairo, Egypt
[4]Landscape Architecture Department, Faculty of Architecture & Planning, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
*Corresponding Author: Mahmoud Ragab. Email: mragab@kau.edu.sa
Received: 30 December 2021; Accepted: 22 February 2022

Abstract: Due to the advanced development in the multimedia-on-demand traffic in different forms of audio, video, and images, has extremely moved on the vision of the Internet of Things (IoT) from scalar to Internet of Multimedia Things (IoMT). Since Unmanned Aerial Vehicles (UAVs) generates a massive quantity of the multimedia data, it becomes a part of IoMT, which are commonly employed in diverse application areas, especially for capturing remote sensing (RS) images. At the same time, the interpretation of the captured RS image also plays a crucial issue, which can be addressed by the multi-label classification and Computational Linguistics based image captioning techniques. To achieve this, this paper presents an efficient low complexity encoding technique with multi-label classification and image captioning for UAV based RS images. The presented model primarily involves the low complexity encoder using the Neighborhood Correlation Sequence (NCS) with a burrows wheeler transform (BWT) technique called LCE-BWT for encoding the RS images captured by the UAV. The application of NCS greatly reduces the computation complexity and requires fewer resources for image transmission. Secondly, deep learning (DL) based shallow convolutional neural network for RS image classification (SCNN-RSIC) technique is presented to determine the multiple class labels of the RS image, shows the novelty of the work. Finally, the Computational Linguistics based Bidirectional Encoder Representations from Transformers (BERT) technique is applied for image captioning, to provide a proficient textual description of the RS image. The performance of the presented technique is tested using the UCM dataset. The simulation outcome implied that the presented model has obtained effective compression performance, reconstructed image quality, classification results, and image captioning outcome.

Keywords: Image captioning; unmanned aerial vehicle; low complexity encoder; remote sensing images; image classification

1 Introduction

The Internet of Multimedia Things (IoMT) is an integration of protocols, interfaces, and multimedia-related data, allows to utilize advanced services and applications depending upon the human-to-device and device-to-device interaction in physical and virtual settings. The faster production in the multimedia-on-demand traffic has moved the vision of the Internet of Things (IoT) from scalar to IoMT. It finds useful in several applications like real time content delivery, online gaming, and video conferencing on the global Internet. On the other hand, Computational Linguistics is a multidisciplinary research field linked to the computation of languages by computers. At present times, the involvement of IoMT and computational linguistics has gained rising attention. Unmanned Aerial Vehicles (UAV) or also named drones generates massive amount of multimedia data, and becomes a part of IoMT. It can be applied under various applications that result in enhanced efficiency in the consumer market. Massive studies have been are extremely concentrated on dealing with communication issues related to UAV and the way to recover the crisis. A drone reaches the complex region and collects data as it lacks external infrastructure. As a result, drones are prominently applied for tedious processes like disaster rescue, observation, transportation in diverse types of applications like agriculture, forestry, climatic prediction, security, and so on. Firstly, drones have been applied autonomously; and in recent times, various synchronized drones are used in operating a complicated process jointly. At this point, drone communication is one of the significant objectives to be computed. Hence, it is essential to learn about different aspects of UAV communication. Besides, various classes of wireless channels as well as network protocols have been applied in drone transmission. Thus, the transmission module is utilized for the UAV system is relied on these applications. For instance, in case of external transmission, it is monitored that sight point-to-point communication connection among drones and a tool could be employed with no breakage.

Traditionally, found different communication as well as mission control methods, for multi-drone schemes, and the classifying networks namely, Decentralized and Centralized. In case of time-sensitive missions, centralized schemes are facilitated in a better way. However, combinations of these methods provide optimal outcomes in which drones are centrally computed and understand from one another. Bluetooth, WiFi, acoustic, ZigBee, and cellular methodologies have been examining for UAV communication module. As a result, the decision of the communication method is operated by regarding the parameter includes efficiency, bandwidth, range, energy demands, cost, compatibility, and payload weight. Diverse models for a drone system with various performances [1] like sensing, coordination, communication, and networking are available. The helpful recommendations are provided such that drones are placed within newly developing huge networks like upcoming cellular systems. Recent wireless networking methods [2] are not applicable to higher mobility of UAVs and enhanced signal frequencies. The Doppler Effect in relative speeds as well as antenna direction related to UAVs results in maximum packet losses. Selection of proper communication models considers antenna device, accuracy, sum rate, and resource managing platforms is recommended. Data transmission is one of the eminent factors a system, and proper routing protocol has to be applied. In case of single UAV, networking is a significant attribute. Fig. 1 shows the structure of UAV.

Drones were combined within the Wireless Sensor Network (WSN), vehicular transmission network, and mobile transmission system for extending the applications of IoT. Artificial intelligence (AI), navigation principles, and cryptography were combined within UAV communication modules by many developers for retaining effective, scalable, and lower latency communications among nodes of UAV system. Therefore, it is essential for considering power efficiency and robustness of drones for consistent for secure communication. Also, it experiences problems like insufficient power and processing resources. Developers have provided insights for optimizing solutions for the applied issues.

Alternatively, the communication failure happens because of aerial network jamming. This interference is considered a severe problem. Networks of UAVs are applied for immediate communication structure as well as surveillance, as presented in [3]. In order to develop an appropriate better UAV communication system, communication and protocols are extremely significant. Diverse technologies are recommended by researchers where crucial aspects like antenna development, network structure, and resource management environments, have been assumed.

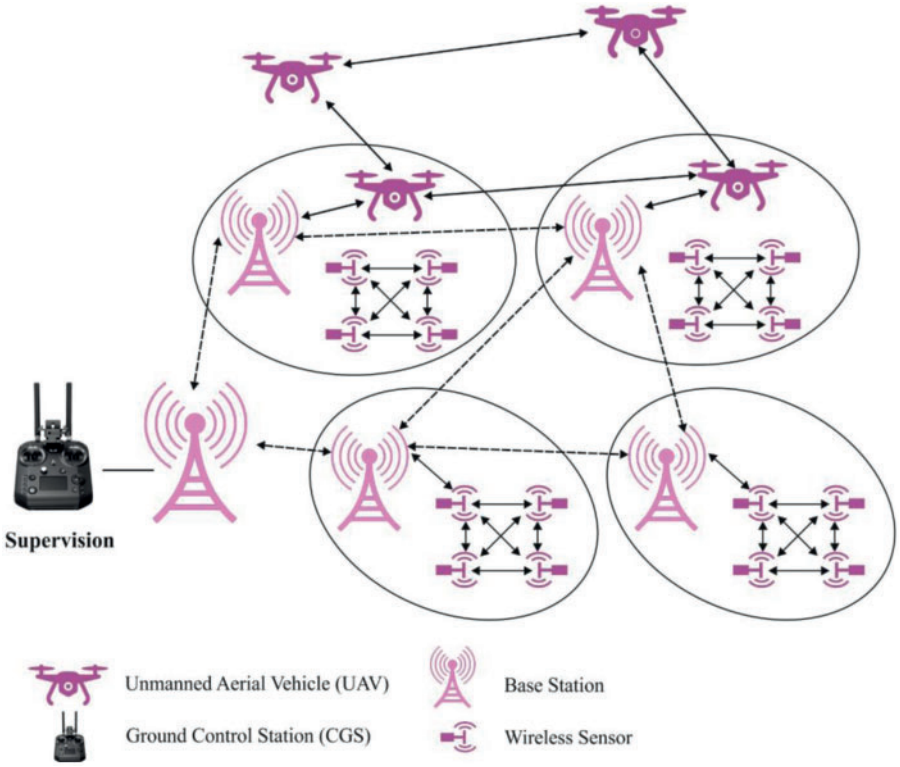


Figure 1: The structure of UAVs

Based on the UAV ecosystem, Deep Learning (DL) is defined as a mainstream tool for massive domains like automated navigation of UAV, object or vehicle observation, and object prediction under real-world limitations. Few works integrate object prediction and monitoring or execute automated aerial reconnaissance operations. Therefore, minimum works consider the advantages from Regions of Interest (ROI), at time in real-world scenarios that is a primary step towards the visual concentration. It is trusted that saliency and dictionary of biases would activate the improvement of present solutions, even under real-world applications. On the other hand, image captioning is a challenging operation for generating a descriptive sentence inside an image automatically, has accomplished a better concentration as eminent interdisciplinary research in Computer Vision (CV) and Natural Language Processing (NLP). Essential applications like guiding visually-impaired users for learning image content or enhancing Image Retrieval (IR) supremacy by identifying salient content. For humans, it is highly simple to attain and it is complex for system as it does not examine certain objects and correlation among images, however, it requires a combination of previous units in appropriate sentences.

This paper introduces a novel low complexity encoding technique with multi-label classification and computational linguistics based image captioning for UAV-based remote sensing (RS) images. The presented model comprises three main processes such as low complexity image encoder, image classification, and image captioning. Initially, the low complexity encoder using Neighborhood Correlation Sequence (NCS) with burrows wheeler transform (BWT) technique called LCE-BWT is employed for encoding the RS images captured by the UAV. Next, the deep learning (DL) based shallow CNN for RS image classification (SCNN-RSIC) approach is developed for determining the distinct class labels of the RS image. Besides, the computational linguistics based Bidirectional Encoder Representations from Transformers (BERT) technique is applied for image captioning, to provide a proficient textual description of the RS image. An extensive experimental analysis is carried out to ensure the effectual outcome of the presented technique is tested using the UCM dataset.

The rest of the paper is organized as follows. Section 2 offers the related works, Section 3 provides the proposed model, Section 4 elaborates the result analysis, and Section 5 draws conclusion.

2 Related Works

Ullah et al. [4] presented a unified structure of UAV as well as Body Area Networks (BANs). Here, it performs the data collection and computes the health data in actual connection with UAV and BAN. In [5], developers have been involved in classification of myocardial infarction as well as atrial fibrillation for examining a signal pattern for various Heart Diseases (HD) by using DL methods. In [6], researchers presented a zero-watermarking technology for securing patients' identity while transmission of clinical data through IoT. In this application, the encrypted key is incorporated within patient's identity image. Additionally, defines the facility of optimal clinical services and requirements for future-generation of telemedicine and telehealth networks. It offers superfast broadband, ultralow delay over different telemedicine to withstand remote healthcare data as well as clinical diagnostics. Hence, it is not real-time for patient home care facilities.

A researcher in [7] defines the requirements of multiple user video streaming along with quality of service (QoS) over wireless systems. Hence, it offers a better solution for video streaming across multi-hop, multi-channel, multi-radio wireless systems. A distribution scheduling model for minimizing video distortion for accomplishing required fairness intensity. Hence, it avoids video quality decomposition because of inherent transmission failures. Moreover, in [8] various QoS technologies and evaluate various integrations for enhancing the supremacy of video stream communicated through unstable system. Therefore, re-encoding or transcoding is essential for decoding various tools.

Cui et al. [9] presented a fast-mode decision mechanism for expediting scalable video coding (SVC) and make use of the correlation among rate-distortion (RD) as well as statistical mode decision of enhancement layer. Paluri et al. [10] proposed a minimum difficulty and delay generalized method to predict the loss of coded video parts for prioritization. It has used this method for the purpose of loss prediction for cumulative mean squared error which is because of drop in coded slices. It employs unequal safety for a slice. The major limitation of this approach is that it is not applicable to use local video content features. Buhari et al. [11] introduced a human visual system-related watermarking scheme with limited complication for acquiring the advantages of texture features.

Koziri et al. [12] found the challenges for reducing load inequalities between diverse threads in slice-based parallelization for proper slice sizing. It uses the collection of pictures as it can be significant for minimum delay situations; however, the drawback is that it enhances the processing difficulty. Santos et al. [13] applied bidirectional prediction support for enhancing the effectiveness of rate predictor's lossless encoding device. Ali et al. [14] developed a sub-partition for extant portions

in the H.264/AVC codec for coding prioritized data. One of the drawbacks is that the computational complexity is enhanced. Grois and Hadar [15] projected coding for ROI, which is decided from pre-encoded scalable video bitstream and adaptive settings activate the extraction of required ROI by position, size, and resolution. It offers an effective approach for meeting satisfy the demand for dissimilar end-user devices. An important limitation of this scheme is the maximized intricacy under various inherent to erroneous platforms.

3 The Proposed Model

The workflow involved in the presented model is illustrated in Fig. 2. As depicted in figure, the UAVs initially capture the images and they are converted into a set of frames. Next, the preprocessing occurs to enhance the quality of the image. Next, the LCE-BWT technique performs encoding of the frames using NCS and BWT by exploiting the relationship among the adjacent pixels. Followed by, the encoded frames are transmitted to the decoder side where the classification and captioning processes were carried out. At the decoder, the LCE-BWT technique is again employed in a reversible way for obtaining the decoded images with no loss of quality. Consequently, the SCNN-RSIC technique is employed for describing the multiple class labels of the RS images. Finally, the BERT technique is utilized to derive the description for the classified RS image.

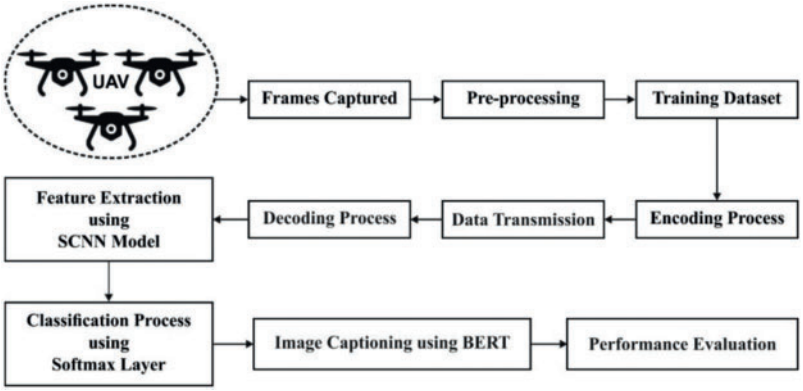


Figure 2: The overall process of proposed method

3.1 Image Encoding Process Using LCE-BWT Model

At this stage, the captured RS images are encoded by the use of NIS technique and then the codewords are compressed using BWT. The LCE-BWT is a scalable and minimal complex image compression model, especially for resource-constrained UAVs. It carried out the compression operation on-the-fly and dynamically extends with modifying source. It is operated in 2 phases namely, Bit reduction with the help of NCS method and encoded by using BWT.

The NCS method produces an optimum codeword for all pixels measure relied on "bit traversal operation under the application of 0's and 1's" [16]. By applying 0's and 1's based traversal, NCS method releases 2 code-words for single pixel value, and codeword with low value of bits are decided as best codeword. Once the code-word generation is completed, codewords undergo encoding by the application of BWT for the compression of image. Consider G as an input image with pixels $\phi_{m,n}$ which is denoted by 2D array as presented in Eq. (1).

$$G = \begin{bmatrix} \phi_{0,0} & \phi_{0,1} & \cdots & \phi_{0,n-1} \\ \phi_{1,0} & \phi_{1,1} & \cdots & \phi_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1,0} & \phi_{m-1,1} & \cdots & \phi_{m-1,n-1} \end{bmatrix} \tag{1}$$

Here m and n denotes the height and width of an input image G. A measure of $m \times n = N$ offers overall resolution of an image G, $\phi_{m,n}$ implies the pixel suited in m^{th} row as well as n^{th} column of grayscale image G. Finally, count of bits required for saving the best code-words Opt_{size} of input image is processed as

$$Opt_{size} = \sum_{i=1}^{N} NCS_{opt}(i) + \text{control bits} \tag{2}$$

where NCS_{opt} refers to the No. of bits in a codeword. Furthermore, NCS models require excess control bits for the best bit count in compressed information. Especially, the bit count needed for saving a pixel from an image is computed as,

$$NCS_{ch\_av} = \frac{Opt_{size}}{N}, \qquad 0 \le NCS_{ch_{av}} \le 5 \tag{3}$$

As the RGB measures of grayscale image have to be similar, it is sufficient for encoding the RGB values for all individual pixels. In particular, it is pointed that a pixel value ranged from [0, 255] and 8 bits are essential for representing a pixel measure. Therefore, the NCS model is operated from 0 bit and 5 bits for implying feasible pixel scores. Furthermore, the compression task is operated using best code-words and encoded under the application of BWT and produces a compressed file.

3.1.1 Optimal Codeword Generation Using NCS Algorithm

The entire workflow of NCS method is depicted here [16]. At the initial stage, the NCS approach reads pixel measures from actual image and transforms a pixel value as corresponding binary format [17]. Then, a bit traversal operation is processed on the basis of 0's and 1's. A bit traversal is initialized with bit of binary sequence and finds a value of initial bit. If the initial bit is 0, then 0's related traversal is operated and saves control bit such as 00, else 01. Under the application of initial bit as reference bit, traversal operation is initialized from upcoming bit seeks for application of 0's in a binary sequence. Based on the value of 0's, equivalent positions (p) are saved as part of best codeword $(00 - p_1)$.

The aforementioned model is repeated till 0's in a series of found, and locations are maximized respectively $(00 - p_1, p_2, \ldots)$. Once 0's relied traversal is finished, after that 1's relied traversal is implemented. Likewise, 0's related traversal and 1's based traversal seeks for the existence of bits and saves the concerned location as a codeword. However, 1's related traversal attains 1's traversal while 0's related traversal reaches the existence of 0's in a binary series. For each pixel value, 2 codewords are produced on the basis of 0's and 1's related traversal. Followed by, a codeword with lower count of bits is selected as best codeword. Consequently, final codewords of pixel measures are combined with control bits for generating a compressed file. Then, codewords produced for the pixel measures ranged from 0 and 255 s.

3.1.2 Codeword Encoding Using BWT Approach

Once the codewords are produced by the NCS technique, they are again encoded using BWT technique to further reduce the file size. BWT is defined as a reversible transform that clusters same units by organizing the input data. The final outcome of BWT is comprised of similar elements of

data which are followed in various iterations. It makes simple data compression significantly [18]. The final data from this phase is composed similar components of data repeated in multiple iterations. It enables the simple compression of data in which runs of elements are projected. Therefore, run-length encoding (RLE) and move-to-front (MTF) transform could be used for encoding such runs and data compression is carried out.

$$bwt\,(s) = \cup_{w \in A^k} \pi_w\,(w_s) \tag{4}$$

Whereas $A = \{a_1, \ldots, a_n\}$ denotes an alphabet and w indicates a word of length k. $\pi_w\,(w_s)$ represents the permutation of the string w_s. Zeroth-order entropy is not applicable in changing the simulation of permuting a string and the root cause of BWT in data compression is for all substring w of s, where it follows w in s are connected within bwt(s).

3.2 SCNN-RSIC Model for Multi-Label Image Classification

Deep convolution neural network (DCNN) includes MobilenetV1 and MobilenetV2 are composed of massive layers and necessities maximum time duration. With the aim of eliminating the above problem, a shallow CNN approach is developed with some layers and tiny convolution kernels size. Then, SCNN with batch normalization method is comprised of a fully connected (FC) layer, 2-layer conv. layers, softmax layer, as well as 2-layer max -pooling layers. The structure of the SCNN is illustrated in Fig. 3. Primarily, the shallow data attributes are extracted using 3×3 convolutions along with filters. For CNN model, batch normalization (BN) is used for normalizing a feature map attained after a convolution and thus an input values of activation function comes under the range sensitive for input for mitigating the possibility of diminishing. For enhancing the accuracy as well as nonlinear expression of model, a convolutional layer applied by BN and Rectified linear unit (ReLu). Afterward, 2×2 max-pooling has been employed for reducing the data dimension as well as processing complexity.

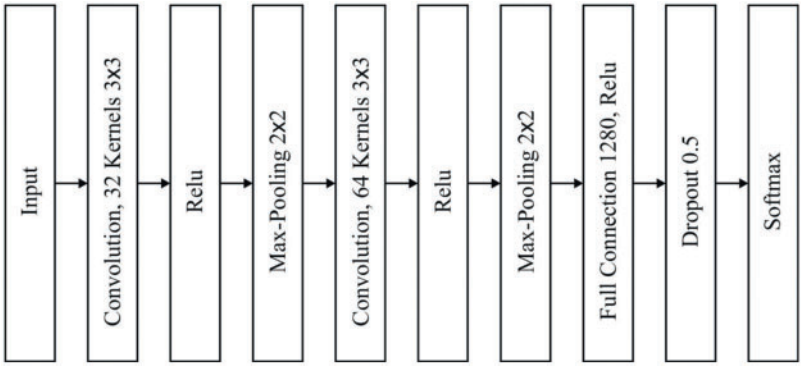


Figure 3: The layers in SCNN

Afterward, 3×3 convolutions along with sixty-four filters for extracting deep data parameters. Next, second convolution layers are composed of Relu and BN which enhanced accuracy and nonlinearity of SCNN technology. The strategies applied in 2×2 max−pooling layers again limit the data dimension as well as computation difficulty [19]. Besides, FC layers with 1280 neurons are applied for combining the features of a topmost layer. Therefore, in FC layer, Relu is applied for enhancing the nonlinear expression capability of technology. Hence, dropout is established for limiting over-fitting and to maximize the generalization potential of technology. Lastly, softmax output layer is applied

for accomplishing multi-classification. It needs tiny processing, storage space, and some iterations, securing valid time resources.

The SCNN scheme has various modalities with the max-pooling layer, input layer, conv. layer, Relu activation function FC layer, softmax output layer, as well dropout principles, as.

- Input layer: Size of an input image is 28×28 for a channel image.
- Convolutional layer: it is an essential layer from CNN, which efficiently extracts the data features. Various convolution kernels obtain distinct data characteristics. Some of the convolution kernels have better potential for extracting optimal features. Hence, SCNN has 2 3×3 conv. layers with filters, correspondingly.
- FC layer: A neuron node of FC layer is linked for a neural node of upper layer, and a neuron node of similar layer is disconnected. Here, SCNN method combines the features of front layer by FC layer with numerous neurons. The FC layer is $1 \times 1 \times 3136$ convolution task, where convolution kernels size is homogeneous with final characteristic size of a former layer.
- Max Z-pooling layer: Once the features are extracted from convolution, a pooling layer has been employed for reducing data redundancy by using down-sampling on obtained features. Therefore, SCNN applies 2 max-pooling layers for reducing data dimension as well as processing complexity whereas the applicable feature extracted remains the same. Therefore, pooling layers does not limit the data repetition and crisis of over-fitting; however, it enhances the training efficiency.
- Relu: it is selected as non-linear activation function of SCNN method. Relu learns from data and map the complex function of input to output and resolve the non-linear issues, and makes further effective which is demonstrated as:

$$f(x) = max(0, x) \tag{5}$$

- Relu is composed of the feature of sparse activation, which reduces over-fitting to a greater amount and enhances the generalization capability of a method.
- Dropout: In DL, when the training parameters are maximum and minimum input data, it results in higher training accuracy and minimum testing accuracy like overfitting. SCNN applies dropout mechanism to randomly remove the neurons of special possibility on FC layer to mitigate over-fitting and stimulates the network training.

For classification purposes, Softmax output layer is used. The frequently applied SCNN applies softmax output layer for accomplishing multi-classification which is illustrated as:

$$softmax\ (y)_i = \frac{e^{y_i}}{\sum_{j=i}^{n} e^{y_i}} \tag{6}$$

where n means a count of output layer nodes, which corresponds to count of classes of certain classifier, and y_i implies a result of *ith* node of resultant layer. Therefore, simulation outcome of this method is transformed as probability distribution using softmax function.

3.3 BERT Based Image Captioning Process

Image captioning is described as a tedious process on intersecting CV and NLP that involves producing small that refers to an image encoder-decoder structures. The tasks of producing captures are sequential. Higher images are organized as "visual language" in which image captioning task is considered as machine translates an operation from a "visual language" to human language. Followed by, massive technologies have showcased as machine translation problems for solving the

image captioning problems of BERT from developers [20]. The strategy behind BERT learning is to learn a model to predict imputed words in sentences. Hence, some parts of words are interchanged with a special token (MASK) and it is applied for predicting words from this context. Afterward, it is to improve the model quality and to gain NN for learning the association among sentences simultaneously whether a single phrase is a logical continuation. Also, BERT is fine-tuned and used for special tasks from NLP. Such features have activated almost part of up-to-date NLP methods for today.

4 Experimental Validation

The performance of the presented system is evaluated by utilizing the benchmark UCM dataset [21,22]. It comprises a set of 100 images under 21 class labels with an image size of 256∗256. The different class labels exist in the dataset are Airplane, Agricultural, Beach, Chaparral, Buildings, Baseball Diamond, Golf Course, Harbor, Intersection, Dense Residential Forest, Freeway, Mobile HomePark, Medium Residential, Parking Lot, Overpass, Runway, Sparse Residential, River, Tennis Court, and Storage Tanks. The image is extracted manually from the USGS National Map Urban Area Imagery gathered to several urban regions over the country. A pixel resolution of the public domain images is around 1 foot. Fig. 4 depicts the sample image from the UCM data set.

Figure 4: The sample images from UCM dataset

Tab. 1 offers detailed results analysis of the presented LCE-BWT system interms of CR and PSNR [23–26]. Fig. 5 illustrates the CR analysis of the presented LCE-BWT model with the existing models under various images. The figure depicted that the LCE-BWT model has resulted in superior compression performance by obtaining a better CR. At the same time, the Deflate technique has shown inferior compression efficiency over all the other methods. For instance, on the applied image 007, the presented LCE-BWT technique has achieved an effective CR of 0.214 whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the CR of 0.342, 0.413, and 0.470 respectively.

Similarly, on the applied image 203, the presented LCE-BWT technique has achieved an effective CR of 0.364 while the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the CR of 0.398, 0.429, and 0.466 respectively. At the same time, on the applied image 457, the presented LCE-BWT method has achieved an effective CR of 0.173 whereas the JPEG, LZE, and Deflate models

have demonstrated its inefficiency with the CR of 0.372, 0.295, and 0.429 respectively. Likewise, on the applied image 692, the presented LCE-BWT technique has achieved an effective CR of 0.238 whereas the JPEG, LZE, and Deflate models have demonstrated its poor CR of 0.297, 0.376, and 0.431 correspondingly. Eventually, on the applied image 831, the presented LCE-BWT technique has achieved an effective CR of 0.219 whereas the JPEG, LZE, and Deflate methods have demonstrated its inefficiency with the CR of 0.318, 0.392, and 0.465 respectively.

Table 1: Comparison of existing methods with the proposed NCS method in terms of CR and PSNR

Images	Compression Ratio (CR)				PSNR (dB)			
	LCE-BWT	JPEG	LZW	Deflate	LCE-BWT	JPEG	LZW	Deflate
Image 007	0.214	0.342	0.413	0.470	56.78	42.18	44.50	43.22
Image 203	0.364	0.398	0.429	0.466	54.69	40.59	45.67	43.80
Image 457	0.173	0.372	0.395	0.429	59.86	39.87	43.85	41.21
Image 692	0.238	0.297	0.376	0.431	53.54	40.11	43.52	42.08
Image 831	0.219	0.318	0.392	0.465	54.19	42.49	43.95	42.51

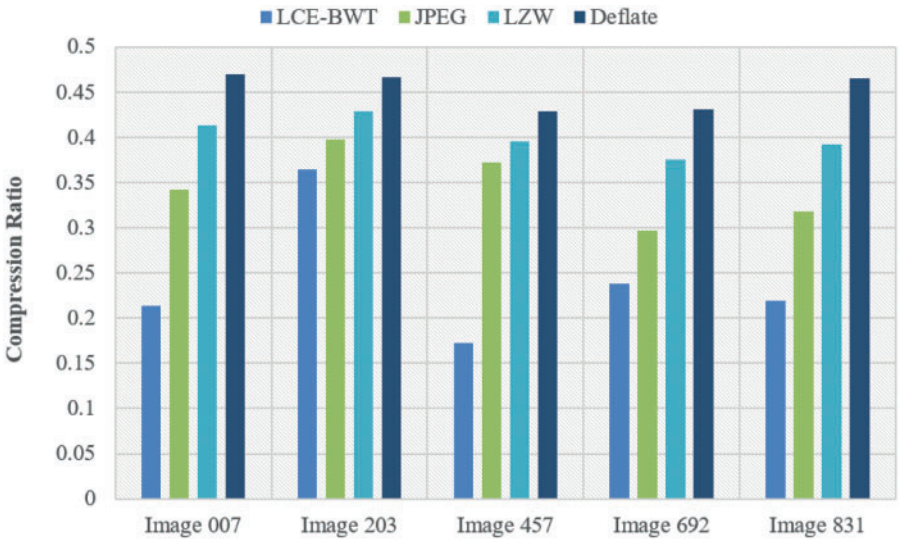


Figure 5: The CR analysis of LCE-BWT model

Fig. 6 depicted the PSNR analysis of the presented LCE-BWT system with the existing methods under various images. The figure depicted that the LCE-BWT model has resulted to superior compression function by obtaining a better PSNR. At the same time, the LZW technique has shown inferior compression efficiency over all the other methods. For instance, on the applied image 007, the presented LCE-BWT technique has achieved an effective PSNR of 56.78 dB whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the PSNR of 42.18, 44.5, and 43.22 dB respectively. Likewise, on the applied image 203, the presented LCE-BWT technique has achieved an effective PSNR of 54.69 dB whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the PSNR of 40.59, 45.67, and 43.8 dB correspondingly. At the same time, on the

applied image 457, the presented LCE-BWT technique has achieved an effective PSNR of 54.86 dB whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the PSNR of 39.87, 43.85, and 41.21 dB respectively. In line with, on the applied image 692, the presented LCE-BWT technique has accomplished an effective PSNR of 53.54 dB whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the PSNR of 40.11, 43.52, and 42.08 dB respectively. Followed by, on the applied image 831, the presented LCE-BWT technique has achieved an effective PSNR of 54.19 whereas the JPEG, LZE, and Deflate models have demonstrated its inefficiency with the PSNR of 42.49, 43.95, and 42.51 dB correspondingly.

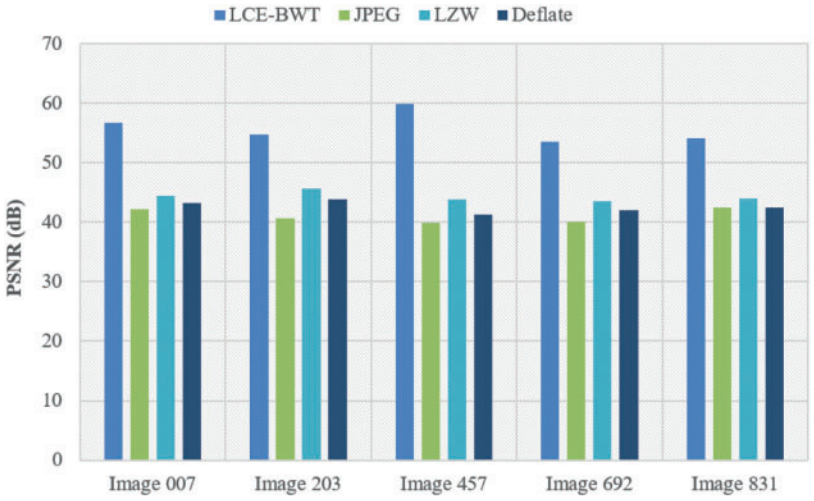


Figure 6: The PSNR analysis of LCE-BWT model

Fig. 7 exhibits the accuracy analysis of the presented SCNN-RSIC model over the other existing methods in the classification of RS images. The table value showcased that the PlacesNet system has accomplished worse RS classification with the least accuracy of 0.914. At the same time, the VFF-VD19 model shows a certainly better RS classification with an accuracy of 0.941. Simultaneously, the AlexNet, CaffeNet, and VGG-F approches have demonstrated same and moderate accuracy of 0.944. Besides, the VGG-M method has tried to portray manageable RS classification outcomes with an accuracy of 0.945 whereas the VGG-S system has reached to even better accuracy of 0.946. However, the presented SCNN-RSIC model has obtained an effective outcome with a maximum accuracy of 0.968.

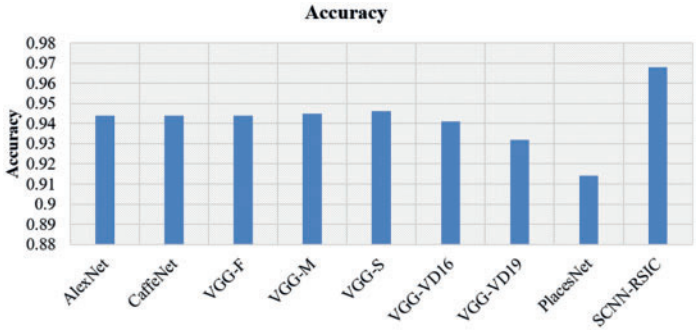


Figure 7: The accuracy analysis of proposed SCNN-RSIC methods with various CNN models

Fig. 8 depicts the result analysis of the presented SCNN-RSIC model over the other existing methods in the classification of RS images interms of F-score. The table values showcased that the VGGNet system has attained worse RS classification with minimal F-score of 0.785. At the same time, the VGG-RBFNN and CA-VGG-LSTM models have offered a certainly better RS classification with the F-score of 0.788 and 0.796. Simultaneously, the ResNet-50, CA-VGG-BiLSTM, and ResNet-RBFNN models have demonstrated moderate and closer F-score of 0.797, 0.798, and 0.806. Besides, the GoogLeNet and CA-ResNet-LSTM methods have tried to portray manageable RS classification outcome with the F-score of 0.807 and 0.814. In line with, the GoogLeNet-RBFNN and CA-ResNet-BiLSTM models have outperformed slightly better and similar results with F-score of 0.815 whereas the CA-GoogLeNet-LSTM and CA-GoogLeNet-BiLSTM models have reached a moderate and identical F-score of 0.818. However, the presented SCNN-RSIC model has obtained an effective outcome with a maximum F-score of 0.891.

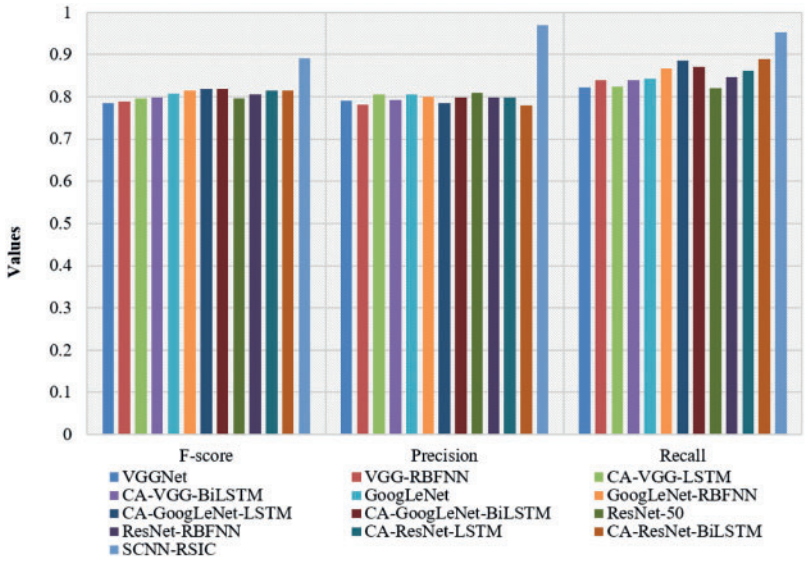


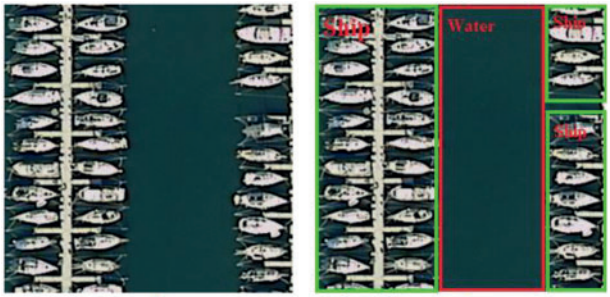
Figure 8: The result analysis of SCNN-RSIC model with different measures

The result analysis of the presented SCNN-RSIC model over the other existing methods in the classification of RS images interms of precision showcased that the CA-ResNet-BiLSTM module has accomplished worse RS classification with the least precision of 0.779. In line with, the CA-GoogLeNet-LSTM method exhibits a moderate outcome with precision of 0.785. At the same time, the CA-GoogLeNet-LSTM, VGGNet, and CA-VGG-BiLSTM models have offered a certainly better RS classification with the precision of 0.785, 0.791, and 0.793. Simultaneously, the CA-GoogLeNet-BiLSTM, CA-ResNet-LSTM, and ResNet-RBFNN models have demonstrated moderate and similar precision of 0.799. Besides, the GoogLeNet-RBFNN and GoogLeNet models have tried to portray manageable RS classification outcome with the precision of 0.800 and 0.805 while the ResNet-50 model has reached to an even better precison of 0.809. However, the presented SCNN-RSIC model has obtained an effective outcome with the maximum precision of 0.969.

The result analysis of the presented SCNN-RSIC model over the other existing methods in the categorization of RS images interms of recall depicted that the ResNet-50 model has accomplished worse RS classification with the least recall of 0.82. At the same time, the VGGNet and CA-VGG-LSTM models have offered a certainly better RS classification with the recall of 0.823

and 0.825. Simultaneously, the VGG-RBFNN, CA-VGG-BiLSTM, and GoogLeNet methods have demonstrated moderate and closer recall of 0.839, 0.84, and 0.843. Besides, the ResNet-RBFNN and CA-ResNet-LSTM models have tried to portray manageable RS classification outcome with the recall of 0.846 and 0.861. In line with, the GoogLeNet-RBFNN and CA-GoogLeNet-BiLSTM models have outperformed slightly better results with recall of 0.868 and 0.871 whereas the CA-GoogLeNet-LSTM and CA-ResNet-BiLSTM models have reached an even better recall of 0.886 and 0.89. However, the projected SCNN-RSIC model has obtained an effective outcome with a maximum recall of 0.953.

Fig. 9 illustrates the visualization results analysis of the presented model. Fig. 9a depicts the actual input image and Fig. 9b shows the classified image with distinct class labels 'ship' and 'water'. The presented model has effectually classified the images and the BERT model has generated the image captioning as 'Ships are around the water. By looking into the detailed simulation analysis, it is evident that the presented model has resulted in effective outcomes in the transmission and analysis of the RS images in UAVs.



(a) (b)

Figure 9: The sample visualization results a) Original image, (b) Classified image

5 Conclusion

This paper has introduced a novel low complexity encoding technique with multi-label classification and image captioning for UAV based RS images. Initially, UAV captures the images and it is transformed into frames. Followed by, the pre-processing operation is computed for enhancing the image quality. Then, the LCE-BWT model computes frame encoding with the help of NCS and BWT by applying the correlation between neighboring pixels. Besides, the encoded frames are then forwarded to a decoder side in which classification and captioning tasks are performed. For a decoder, the LCE-BWT approach is applied in a reversible way to gain decoded images without loss of supremacy. As a result, the SCNN-RSIC framework is used for computing the class labels of RS images. Lastly, the BERT model has been applied for deriving description for categorized RS image. The working principle of the projected method undergoes testing under the application of the UCM dataset. The results showcased that the proposed method has attained significant compression performance, reformed image quality, classification outcomes, as well as image captioning results. In future, the proposed model can be validated using large scale real time datases.

Acknowledgement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the Project Number (IFPIP-941-137-1442) and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Funding Statement: This project was supported financially by Institution Fund projects under Grant No. (IFPIP-941-137-1442).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. Yanmaz, S. Yahyanejad, B. Rinner, H. Hellwagner and C. Bettstetter, "Drone networks: Communications, coordination, and sensing," *Ad Hoc Networks*, vol. 68, pp. 1–15, 2018.

[2] M. Asadpour, D. Giustiniano and K. A. Hummel, "From ground to aerial communication: Dissecting wlan 802.11 n for the drones," in *Proc. of the 8th ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization*, Miami Florida USA, pp. 25–32, 2013.

[3] W. S. Jung, J. Yim, Y. B. Ko and S. Singh, "Acods: Adaptive computation offloading for drone surveillance system," in *Ad Hoc Networking Workshop (Med-HocNet), 2017 16th Annual Mediterranean*, Budva, Montenegro, pp. 1–6, 2017.

[4] S. Ullah, K. I. Kim, K. H. Kim, M. Imran, P. Khan *et al.,* "UAV-enabled healthcare architecture: Issues and challenges," *Future Generation Computer Systems*, vol. 97, pp. 425–432, 2019.

[5] U. Iqbal, T. Y. Wah, M. H. u. Rehman, G. Mujtaba, M. Imran *et al.,* "Deep deterministic learning for pattern recognition of different cardiac diseases through the internet of medical things," *Journal of Medical Systems*, vol. 42, no. 12, pp. 252–265, 2018.

[6] Z. Ali, M. Imran, M. Alsulaiman, M. Shoaib and S. Ullah, "Chaos-based robust method of zero-watermarking for medical signals," *Future Generation Computer Systems*, vol. 88, pp. 400–412, 2018.

[7] L. Zhou, X. Wang, W. Tu, G. M. Muntean and B. Geller, "Distributed scheduling scheme for video streaming over multi-channel multi-radio multi-hop wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 409–419, 2010.

[8] A. Ziviani, B. E. Wolfinger, J. F. de Rezende, O. C. M. B. Duarte and S. Fdida, "Joint adoption of QoS schemes for MPEG streams," *Multimedia Tools and Applications*, vol. 26, no. 1, pp. 59–80, 2005.

[9] Y. Cui, W. Ren and Z. Deng, "Fast mode decision for HD scalable video coding via statistical content analysis," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 8, pp. 1–19, 2018.

[10] S. Paluri, K. K. R. Kambhatla, B. A. Bailey, P. C. Cosman, J. D. Matyjas *et al.,* "A low complexity model for predicting slice loss distortion for prioritizing H.264/AVC video," *Multimedia Tools and Applications*, vol. 75, no. 2, pp. 961–985, 2014.

[11] A. M. Buhari, H. C. Ling, V. M. Baskaran and K. Wong, "Low complexity watermarking scheme for scalable video coding," in *2016 IEEE Int. Conf. on Consumer Electronics-Taiwan (ICCE-TW)*, Nantou, Taiwan, pp. 5–6, 2016.

[12] M. Koziri, P. Papadopoulos, N. Tziritas, A. N. Dadaliaris, T. Loukopoulos *et al.,* "Slice-based parallelization in HEVC encoding: Realizing the potential through efficient load balancing," in *2016 IEEE 18th Int. Workshop on Multimedia Signal Processing (MMSP)*, Montreal, QC, Canada, pp. 1–6, 2016.

[13] J. M. Santos, A. F. R. Guarda, L. A. da Silva Cruz, N. M. M. Rodrigues and S. M. M. Faria, "Compression of medical images using MRP with bidirectional prediction and histogram packing," in *2016 Picture Coding Symp. (PCS)*, Nuremberg, Germany, pp. 1–5, 2016.

[14] I. Ali, S. Moiron, M. Fleury and M. Ghanbari, "Data partitioning technique for improved video prioritization," *Computers*, vol. 6, no. 3, pp. 23, 2017.

[15] D. Grois and O. Hadar, "Efficient region-of-interest scalable video coding with adaptive bit-rate control," *Advances in Multimedia*, vol. 2013, pp. 1–17, 2013.

[16] J. Uthayakumar, M. Elhoseny and K. Shankar, "Highly reliable and low-complexity image compression scheme using neighborhood correlation sequence algorithm in WSN," *IEEE Transactions on Reliability*, vol. 69, no. 4, pp. 1398–1423, 2020.

[17] J. Uthayakumar, T. Vengattaraman and P. Dhavachelvan, "A new lossless neighborhood indexing sequence (NIS) algorithm for data compression in wireless sensor networks," *Ad Hoc Networks*, vol. 83, pp. 149–157, 2019.
[18] A. Khan and A. Khan, "Lossless colour image compression using RCT for bi-level BWCA," *Signal, Image and Video Processing*, vol. 10, no. 3, pp. 601–607, 2016.
[19] F. Lei, X. Liu, Q. Dai and B. W. K. Ling, "Shallow convolutional neural network for image classification," *SN Applied Sciences*, vol. 2, no. 1, pp. 97, 2020.
[20] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[21] http://weegee.vision.ucmerced.edu/datasets/landuse.html.
[22] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. of the 18th SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems-GIS '10*, San Jose, California, pp. 270, 2010.
[23] A. Rajagopal, G. P. Joshi, A. Ramachandran, R. T. Subhalakshmi, M. Khari *et al.,* "A deep learning model based on multi-objective particle swarm optimization for scene classification in unmanned aerial vehicles," *IEEE Access*, vol. 8, pp. 135383–135393, 2020.
[24] Y. Hua, L. Mou and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 188–199, 2019.
[25] K. Qi, Q. Guan, C. Yang, F. Peng, S. Shen *et al.,* "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sensing*, vol. 10, no. 6, pp. 934, 2018.
[26] A. Rajagopal, A. Ramachandran, K. Shankar, M. Khari, S. Jha *et al.,* "Fine-tuned residual network-based features with latent variable support vector machine-based optimal scene classification model for unmanned aerial vehicles," *IEEE Access*, vol. 8, pp. 118396–118404, 2020.