PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Advanced Methods for Content Based Image Retrieval and Scene Classification in JPEG 2000 Compressed Remote Sensing Image Archives

Akshara Preethy Byju

Advisor:	Co-advisor:
Prof. Dr. Lorenzo Bruzzone	Prof. Dr. Begum Demir
Università degli Studi di Trento	Technische Universitat Berlin

January 2020

Abstract

Recent advances in satellite imaging technologies have paved its way to the RS big data era. Efficient storage, management and utilization of massive amounts of data is one of the major challenges faced by the remote sensing (RS) community. To minimize the storage requirements and speed up the transmission rate, RS images are compressed before archiving. Accordingly, developing efficient Content Based Image Retrieval (CBIR) and scene classification techniques to effectively utilize these huge volume of data is one among the most researched areas in RS. With the continual growth in the volume of compressed RS data, the dominant aspect that plays a key role in the development of these techniques is the decompression time required by these images. Existing CBIR and scene classification methods in RS require fully decompressed RS images as input, which is a computationally complex and time consuming task to perform. Among several compression algorithms introduced to RS, JPEG 2000 is the most widely used in operational satellites due to its multiresolution paradigm, scalability and high compression ratio. In light of this, the goal of this thesis is to develop novel methods to achieve image retrieval and scene classification for JPEG 2000 compressed RS image archives.

The first contribution of the thesis addresses the possibility of performing CBIR directly on compressed RS images. The aim of the proposed method is to achieve efficient image characterization and retrieval within the JPEG 2000 compressed domain. The proposed progressive image retrieval approach achieves a coarse to fine image description and retrieval in the partially decoded JPEG 2000 compressed domain. Its aims to reduce the computational time required by the CBIR system for compressed RS image archives.

The second contribution of the thesis concerns the possibility of achieving scene classification for JPEG 2000 compressed RS image archives. Recently, deep learning methods have demonstrated a cutting edge improvement in scene classification performance in large-scale RS image archives. In view of this, the proposed method is based on deep learning and aims to achieve maximum scene classification accuracy with minimal decoding. The proposed approximation approach learns the high-level hierarchical image description in a partially decoded domain thereby avoiding the requirement to fully decode the images from the archive before any scene classification is performed.

Quantitative as well as qualitative experimental results demonstrate the efficiency of the proposed methods, which show significant improvements over state-of-the-art methods.

Keywords: content based image retrieval, deep neural networks (DNNs), scene classification, progressive image retrieval, JPEG 2000, compressed image domain, pyramid match kernel, remote sensing

Acknowledgements

First and foremost, I would like to thank my parents for their unconditional love and support that they have given me throughout all these years. No words would be enough to thank them for their belief, prayers and support that they have given me. A special thanks to all my family members especially paathu and mittu for being very kind and loving me in all respects.

I would like to whole-heartedly thank Prof. Dr. Begum Demir for her continuous support, advice and guidance since the beginning to the end of my PhD. I really appreciate the time, ideas and opportunity that she provided to carry out the research in the best way possible. I really admire her quality of being extremely thoughtful that has motivated me to be a better person in life and I feel I have been particularly lucky to be her student as well as a part of her research group.

I would like to express my deepest gratitude to Prof. Dr. Lorenzo Bruzzone for all the priceless advice that he provided during the duration of my PhD. It is always a pleasant and peaceful experience to look forward to the meetings so as to receive very insightful research directions from him and has motivated me to finish the thesis. I consider this as a great privilege to be his student and to be a part of his research team.

I would like to thank all my friends for making this experience a bit better and happier. My special to thanks to Dr. Aravind Harikumar for being such as amazing friend to give all advice, help and support since the beginning of my PhD. I would like to thank my Indian friends Ashky, Sara, Saifu, Megha and Mahesh for all the laughs and I am extremely glad to meet them through this walk of life. I would also like to thank Zafer, Gencer, Kimya, Alaro, Ruben, Sue, Sophia, Adela and Jian for making my experience a bit extraordinary at RSiM. Last but not the least, I would like to thank all the Remote Sensing Lab (RSLab) members for being one of the most amazing colleagues to me.

Akshara Preethy Byju

Dedicated to my parents

Contents

List of Tables

List of Figures

List of Abbreviations

RS	Remote Sensing
DPCM	Differential Pulse Code Modulation
ADPCM	Adaptive Differential Pulse Code Modulation
SQNR	signal-to-quantization-noise ratio
JPEG	Joint Photographic Experts Group
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
CBIR	Content Based Image Retrieval
DL	deep learning
DNNs	Deep Neural Networks
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
GANs	Generative Adversarial Networks
EBCOT	Entropy Block Coding with Optimized Truncation
ROI	Region-Of-Interest
TCQ	trellis coded quantization
MSB	Most Significant Bit
LSB	Least Significant Bit
ANNs	artificial neural networks
FC	Fully Connected
LBP	Local Binary Pattern
GLCM	Grey level Co-occurrence Matrix

HOG	Histogram of Gradients
SIFT	Scale-Invariant Feature Transform
BoVW	bag-of-visual-words
VLAD	Vector of Locally Aggregated Descriptors
MCNN	multi-scale CNN
HI	Histogram Intersection
KLD	Kullback Leibler Divergence
SVMs	Support Vector Machines
MLCs	Maximum Likelihood Classifiers
SPM	Spatial pyramid matching
SSPM	spatial-spectral pyramid matching
MSP	multiscale pooling
RF	Relevance Feedback
AE	Autoencoders
HSI	hyperspectral images
MSGAN	multiclass spatial-spectral GAN
DCGAN	deep convolutional GANs
LSTM	Long Short-Term Memory
MBH	Maximum Bit Histogram
GGD	Generalized Gaussian Distribution
GMM	Gaussian Mixture Model
PM	pyramid match

List of symbols

$\mathbf{X}$	an archive of N JPEG 2000 compressed-images
X_i	i-th image in the archive $\mathbf{X}$
X_q	query image
X^{rel}	set of relevant images
L	wavelet decomposition levels used in the archive $\mathbf{X}$
$h^l_{X_i}$	horizontal sub-band of an image X_i at l-th wavelet decomposition level
$v^l_{X_i}$	vertical sub-band of an image X_i at l-th wavelet decomposition level
$H^l_{X_i}$	histogram obtained at l-th wavelet decomposition level for image X_i
r	the number of histogram intervals
$\varphi^l_{X_i}(u, v)$	moduli of the horizontal and vertical wavelet sub-band coefficients obtained at sample location (u,v)
m	number of rows in the considered wavelet subband
n	number of columns in the considered wavelet subband
w_L	weight values of the descriptors associated to the coarsest wavelet L-th wavelet resolution

N_m	implicit partial correspondence between any two successive wavelet decomposition levels
T_1	percentage of discarded images at the first level
T_2	percentage of discarded images at the second level
$\mathbf{Q}$	total number of class labels
$\mathbf{Y}$	set of $\mathbf{Q}$ class labels
y_i	class label associated to image X_i
G^L	coarsest approximation, vertical, horizontal and detail wavelet sub-band at level L of image X_i
A^{L-1}	finer approximation, vertical, horizontal and detail wavelet sub-band at level *L-1* of image X_i
$\mathbf{C}$	sparse matrix
S	stride used in the DNN
P	padding used in the DNN
A^{L-1}_{size}	size of the approximated wavelet subband
A^{L}_{size}	size of the coarsest wavelet subband
$\mathcal{L}_{total}$	total loss function
$\mathcal{L}_{classification}$	classification loss function
$\mathcal{L}_{approximation}$	approximation loss function
D^i	Decoded wavelet sub-band at any given wavelet decomposition level i
$\hat{y}_i$	the predicted class label

Chapter 1

Introduction

1.1 Background and Motivation

The acquisition of satellite remote sensing data started in the early 1970's and since then there has been an exponential growth in the development of this technology. Remote Sensing (RS) data provide relevant information on the Earth's surface and are used in wide number of applications such as agriculture, environment monitoring, disaster management, meteorology, oceanography and urban planning. In the recent years, the remarkable progress in the development of imaging sensors and satellite missions, which has contributed to a massive growth in the volume of acquired RS data. Developments in active and passive sensor technologies have contributed to the so called *RS big data era*. RS big data are characterized by six key properties (Fig. 1.1): *Volume*, *Velocity*, *Variety*, *Veracity*, *Value* and *Visualization* (6V dimensions) [1]. *Volume* indicates the large amount of data that are acquired by the satellites. As an example, European Space Agency's (ESA) Sentinel missions (Sentinel-1, Sentinel-2 and Sentinel-3) alone provide 10 Terabytes (10 TB) of data per day [2]. At the end of 2019, the volume of data that is archived from the Sentinel missions alone is estimated to be more than 12.5 PB [3]. *Variety* refers to the distinct continuous group of data obtained from multisource (e.g., multispectral, hyperspectral, Synthetic Aperture Radar (SAR)) and multi-temporal (time series) images. *Velocity* refers to the speed of generation of the incoming data. For example, Sentinel-2 and PRISMA (hyperspectral) missions have a revisit time of 5 and 7 days, respectively [4; 5]. *Veracity* refers to the reliability and effectiveness of the acquired RS data. The massive amount of data that are acquired every day must be accurate enough to be efficiently utilized by the RS society. Lastly, *Value* and *Visualization* are other two key aspects in RS big data that involve information loss and noise generated from the many satellite missions [1]. In view of these features, developing efficient techniques for storing, managing and effectively utilizing massive volumes of big data is one among the major challenges faced by the RS community.

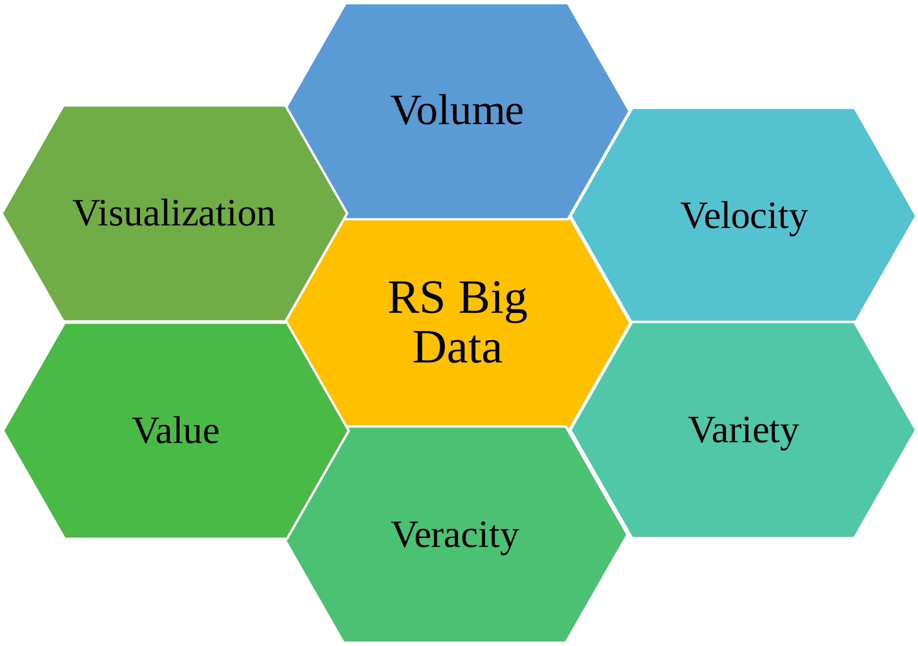


Figure 1.1: RS big data properties.

The availability of large volumes of data from satellite sensors with increased spatial, spectral and radiometric resolution demands more storage space and thus, it is required to compress images before storing them into any archive [6; 7; 8; 9; 10; 11; 12; 13; 14; 15]. Compression algorithms can be categorized as lossy and lossless. Lossless compression techniques reduce the size of the images without degrading their quality whereas lossy compression technique achieves higher compression ratio with a significant quality degradation. In view of this, several compression algorithms were proposed in the RS literature [7]. In the early 1980s, predictive coding compression techniques such as Differential Pulse Code Modulation (DPCM) and Adaptive Differential Pulse Code Modulation (ADPCM) were used in SPOT 1,2,3 satellites [16]. DPCM encodes data by predicting the difference between the reference and previous pixel in a given image. If the correlation between any two adjacent pixels is very small the signal-to-quantization-noise ratio (SQNR) also decreases. Considering the current volume of the archives, the compression ratio achieved by these predictive coding techniques is considerably low. In order to achieve better compression that leads to a better use of the storage space, transform based coding techniques have proposed where the images are converted into different domains. Joint Photographic Experts Group (JPEG) was one among the first transform coding compression techniques developed during the early 1990s [17]. JPEG uses Discrete Cosine Transform (DCT) which is a lossy compression technique that initially subdivides the image into several blocks and then applies DCT to each of these blocks to obtain image blocks with varying frequencies. As DCT is performed on each block of a given image, it generates blocking artifacts which generally occur at the boundaries of each block of the considered image. In addition, JPEG compression algorithm is lossy and does not support lossless compression. To address these limitations, JPEG 2000 image compression standard was proposed [18]. JPEG 2000 uses Discrete Wavelet Transform (DWT), which allows for a multiresolution representation of the images. The inherent multiresolution paradigm within the JPEG 2000 image coding standard achieves higher compression ratio compared to other compression algorithms, scalability and progressive transmission of the images. Most of the operational RS satellites compress their images before storing into the archive (e.g., Sentinel-2 uses JPEG 2000 compression algorithm and PRSIMA uses wavelet based compression module) [19; 20].

In the past decade, several efforts were made to retrieve domain-specific relevant information at a fast rate from massive RS image archives. Traditional image retrieval techniques strongly rely on metadata (such as keywords, tags, location, acquisition date or time) to extract images from the archive. Recent developments demonstrated that also in RS image retrieval can be done on the basis of the 'contents'. Thus, Content Based Image Retrieval (CBIR) gained increasing attention in the RS community [21; 22]. When a user provides a query image, the aim of the CBIR system is to retrieve similar images associated with it. To achieve this, CBIR systems utilize the implicit information within the considered image and archive images. Thus, the performance of a given CBIR system depends mainly on efficient representation of this implicit information (feature descriptors) of images in the archive. A general CBIR system has two steps: i) obtaining efficient feature descriptors; and ii) similarity assessment between the query image and archive images. The existing RS CBIR techniques require fully decompressed (decoded) image as input. Moreover, decompressing each image from a massive archive and applying feature extraction and similarity assessment with respect to the submitted query image is a time demanding and computationally complex task. In view of this, the thesis contributes to the development of RS CBIR methods for JPEG 2000 compressed RS image archives.

In the recent years, deep learning (DL) emerged as one of the major breakthrough in several RS image processing tasks [23; 24]. DL methods have demonstrated significant success in several domains such as image retrieval, change detection and objection detection. Among them, scene classification is of particular interest [25; 26; 27; 28]. Considering the huge volume of data that are stored in image archives, scene classification approaches based on DL approaches have become very popular as they can manage the complexities of handling massive amounts of multi-dimensional data in terms of feature descriptors and assessing similarity. The excellent performance of the DL methods in obtaining highly discriminative feature descriptors automatically through the learning process has become one of the major driving force that led towards their success. To learn effective feature descriptors, DL methods require huge amounts of data that usually are processed by powerful Graphic Processing Units (GPUs). Several DL architectures were introduced to address RS scene classification problems [28; 29; 30; 31]. Deep Neural Networks (DNNs) such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs) have demonstrated impressive classification performance. Among them, CNNs have shown a remarkable ability to obtain feature representations that describe high-level semantic patterns of the considered images. Generally, a DL method can be categorized into a supervised or an unsupervised learning scheme. CNNs, are based on a supervised learning scheme and thus require annotated images for the training of the architecture. On the contrary, in the unsupervised scheme, the DL network automatically clusters the data based on the inherent patterns present in the images. Recently, GANs became popular due to their ability to generate and learn feature descriptors from a limited amount of images in specific RS domains. It is very important to emphasize the fact that DL methods are able to achieve impressive performance at varying imaging conditions (such as scale, translation and rotation). The continuous growth in the RS image archives allows DL architectures to learn much efficient feature descriptors. Moreover, also in this context exiting DL architectures that are used to address RS scene classification problems require full decompression of the images before they are provided as input to the network. This is a strong limitation from the computational point of view and require to explore the possibility of developing novel DL architectures that can learn compressed domain features to perform scene classification.

1.2 Motivation and Novel Contribution of the Thesis

As mentioned in the previous section, existing RS image retrieval and scene classification approaches require full decompression of the RS images. This is a time demanding and computationally complex task to carry out in operational systems working on very large archives. In computer vision and pattern recognition, several works were recently proposed to study the possibility of obtaining feature descriptors in the compressed domain. However, existing RS image retrieval systems do not consider this possibility despite the massive volume of compressed images available in RS image archives. Thus, it is necessary to develop novel approaches to efficiently perform image retrieval without the need for fully decompressing all the images in the archive. On the basis of this analysis, the aim of the thesis is to develop computationally efficient methods to perform image retrieval and scene classification in JPEG 2000 compressed domain. In particular, the main novel contributions of the thesis are as follows:

1. A novel progressive Content Based Image Retrieval (CBIR) system that minimizes the amount of decompression required for all the JPEG 2000 compressed images in the archive.

2. An efficient approximation approach within a DNN framework to accurately characterize the JPEG 2000 compressed domain wavelet sub-band information to achieve scene classification in large-scale compressed RS image archives.

In the subsections below, we briefly describe each of the contributed methods.

Progressive Content Based Image Retrieval Approach

The continuous increase in the volume of compressed data in RS image archives demands developing efficient CBIR systems. Although as mentioned there are several RS CBIR systems in the literature that have demonstrated remarkable performances, they require decompression of all the images to apply image description and similarity analysis, which is a time demanding and computationally complex task [21; 22; 32; 33; 34; 35]. In view of this, the first contribution of the thesis is to develop a novel progressive CBIR system that minimizes the amount of decompression required for the retrieval of images from compressed RS image archives. The system is based on the observation that the decoding paradigm within the JPEG 2000 compression algorithm allows to progressively decode the coarse to fine wavelet sub-band information. Thus, in the proposed approach we exploit the possibility to avoid decompression of all the images in the archive by adopting a coarse-to-fine image characterization and retrieval approach. The approach initially decodes the codestreams associated to the coarsest level wavelet sub-band information, which can be utilized to eliminate a few irrelevant images from the archive. Then, the successive finer wavelet sub-band information associated to the subset of relevant images is decoded. The similarity assessment is carried out by considering the features obtained from both coarser and finer level wavelet sub-bands by adopting a pyramid match kernel. The image description and similarity assessment are iterated until we decode the compressed images to obtain the finest wavelet resolution sub-band. In this way, the proposed approach eliminates irrelevant images during the early stages, thereby reduces the required decoding time and thus speeding up the computational time of the proposed CBIR system. The effectiveness of the proposed method is demonstrated on UCMERCED [36] and AID [37] benchmark archives.

Remote Sensing Image Scene Classification using an Approximation Approach

As mentioned in the previous section, all the existing methods require full decompression of the images before scene classification can be carried out [38; 39; 40; 41; 42]. To address this issue, in the second contribution of the thesis, we propose a novel approximation approach that allows to perform scene classification in the compressed RS image archives. In this method, the finer level wavelet information (which is used in JPEG 2000) is approximated through few transposed deconvolutional layers. The approximated finer resolution wavelet sub-band information is learnt through a series of convolutional layers to achieve scene classification. Through this approximation approach, the requirement to fully decompress all the images is minimized thus speeding up the computational time required to perform scene classification in the compressed RS image archives. The proposed method includes a novel end-to-end trainable DNN architecture that efficiently exploits the finer level wavelet sub-band information obtained using the approximation approach to achieve computationally efficient scene classification in the JPEG 2000 compressed domain. The effectiveness of the proposed method is demonstrated on AID [37] and NWPU-RESISC45 [43] benchmark archives.

1.3 Thesis Organization

This chapter has provided an overview of the current scenario and the motivation of the thesis. It also summarized the existing compression algorithms and the challenges that arises to apply image retrieval and scene classification approaches to compressed RS image archives. The rest of the thesis is organized into five chapters.

Chapter 2 illustrates the JPEG 2000 compression algorithm that is fundamental to understand the approaches proposed in the thesis. It also provides some basics on DNNs (especially CNNs) to give the required background on the DL architecture for the next part of the thesis. Chapter 3 provides an analysis of the state-of-the-art on the existing image retrieval and scene classification approaches in RS. It also presents a review of the existing JPEG 2000 compressed domain works on image retrieval and classification in computer vision and pattern recognition.

Chapter 4 presents the novel image RS CBIR system in JPEG 2000 compressed domain providing in detail the methodology, experimental results and a final discussion.

Chapter 5 describes the proposed novel approach using DNN to perform computationally efficient scene classification in JPEG 2000 compressed image archives.

Finally, chapter 6 draws the conclusion of the thesis along with the possible future research developments.

Chapter 2

Background and Fundamentals

In this chapter first we present the fundamentals of JPEG 2000 compression algorithm and then we provide the background of deep neural networks (DNNs).

Several compression algorithms are introduced in the RS literature (e.g., Differential Pulse Code Modulation (DPCM), Adaptive DPCM, Joint Photographic Experts Group (JPEG), lossy and lossless JPEG and JPEG 2000). In order to compress RS data, earlier predictive coding compression techniques such as DPCM and ADPCM were used in satellites such as SPOT 1,2,3 [44; 45]. DPCM encodes data by predicting the difference between the reference pixel and the previous pixel. Although, using DPCM the inter-band dependency is reduced and compression ratio is improved, with the increasing number of RS data in the archives, these algorithms increase the computational complexity in image retrieval as well as scene classification techniques. However, these techniques cannot be used in archives where the numbers of images are huge and when there is a requirement to obtain higher compression rate. Also, predictive based coding techniques are quite complex when compared to other compression algorithms and extracting features from them for RS data adds further computational overhead [7]. To achieve higher compression ratio, transform based coding techniques were proposed where the images are converted to their frequency domain. Among various transform based coding techniques better compression rates were achieved with Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) [7]. JPEG and JPEG 2000 compression standards have adopted DCT and DWT as their transformation techniques, respectively [17; 46]. Among the aforementioned compression techniques, JPEG 2000 became very popular due to its ability to achieve multiresolution paradigm, scalability and high compression ratio. In this thesis, we focus on developing novel approaches to perform content based image retrieval and scene classification in JPEG 2000 compressed image archives.

In view of this, Section 2.1 discusses the general block scheme of JPEG 2000 compression algorithm. Section 2.2 introduces the key concepts of CNN. These sections presents the necessary information that is required for the understanding of the methods proposed in this thesis.

2.1 Overview of the JPEG 2000 Compression Algorithm

JPEG 2000 image compression standard was developed by the Joint Photographic Experts Group which is a joint committee of members from the International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) [46; 47]. JPEG 2000 is the successor of the JPEG compression standard that uses Discrete Cosine Transform (DCT) which was initially developed in 1992. Figure 2.1 shows the general block scheme of the JPEG 2000 encoder. When a compression algorithm is considered, it can be either lossless (where the image compression is achieved without information loss) or lossy (where the image compression is achieved with data loss). JPEG 2000 standard supports both lossless as well as lossy compression. The encoder of JPEG 2000 compression algorithm consists of three main blocks: i) Discrete Wavelet Transform; ii) Quantization and iii) Entropy Block Coding with Optimized Truncation (EBCOT). The original image can also be decoded (decompressed) from the compressed codestream using three steps: i) Entropy decoding; ii) De-quantization and iii) Inverse Discrete Wavelet Transform (IDWT). The main properties of the JPEG 2000 compression algorithm that makes it successful are as follows:

1. The images are represented using multi-resolution representation that is achieved using DWT used in the compression standard.

2. Both lossy as well as lossless compression are supported with higher compression ratios compared to the previous compression algorithms.

3. The packet structure organization of the wavelet sub-band allows progressive transmission and decoding of the images based on resolution, quality, spectral band and location.

4. Region-Of-Interest (ROI) coding of a given image is allowed.

5. Robustness to the transmission errors are proven.

In the following subsections, the three main blocks of the JPEG 2000 encoder are detailed. For more details regarding the JPEG 2000 compression algorithm, the Reader is referred to [18].

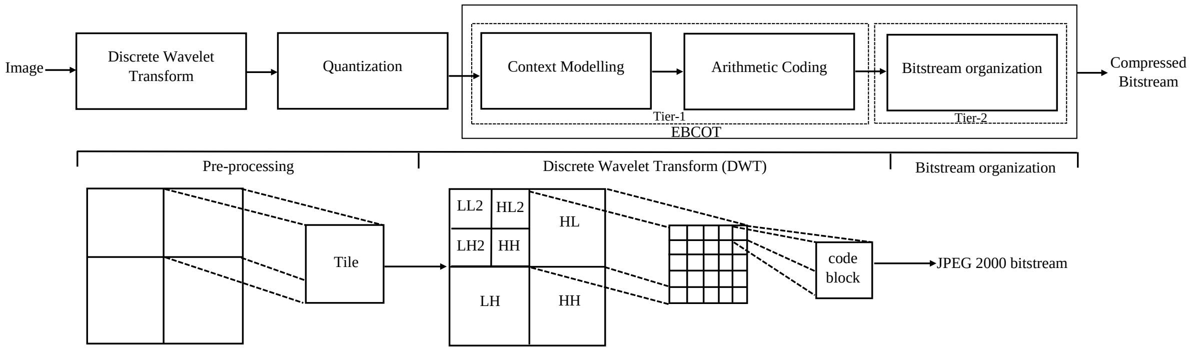


Figure 2.1: General block scheme of the JPEG 2000 compression algorithm.

2.1.1 Discrete Wavelet Transform (DWT)

DWT is one of the important steps in the JPEG 2000 compression algorithm that results in successive dyadic wavelet decomposition (which allows multiresolution analysis) of a given image [18]. This multiresolution image representation form is utilised in several domains such as image compression, noise removal, etc. Wavelets are mathematical functions that decompose a given input image into several frequency components and analyze each component based on the considered scale (resolutions) [48]. Let $\psi(x)$ be a 'mother wavelet' (wavelet function) of a wavelet basis $L_2(\mathbb{R})$. Then, a family of wavelet function $\psi_{s,d}(x)$ is obtained using scaling (s) and dilation (d) of the mother wavelet $\psi(x)$ as:

$$\psi_{s,d}(x) = \sqrt{2^s}\psi(2^s x - d) \tag{2.1}$$

DWT of a given image X_i is obtained by passing it through a series of low-pass and high-pass filters. If l is an impulse response, then the output of the low-pass filter is a convolution of X_i and g which is obtained as:

$$Y_i[n] = (X_i \circledast l)[n] \tag{2.2}$$

where Y_i represents the approximation coefficients. The input image X_i is also passed through a high pass filter h which results in detail coefficients. Fig. 2.2 represents low-pass and high-pass filter bank realization of a given image X_i. Successive dyadic wavelet decomposition applied to each image or *tile* separately transforms an image into one low frequency (approximation coefficients - LL) sub-band and three high frequency sub-bands (detailed coefficients - LH, HL and HH). If there are more than one decomposition level, the lowest (LL) sub-band of the current resolution is further decomposed to the subsequent approximation and detail wavelet resolution sub-bands. The maximum number of wavelet decompositions that can be performed on a given image according to the JPEG 2000 compression standard is 32. The approximation (LL) wavelet coefficients encompass the information of the original input image at a lower wavelet resolution. The detail subbands include the information such as edge, texture, etc of a given image. In the JPEG 2000 compression standard, successive dyadic wavelet decomposition for both lossy and lossless compression is performed using Cohen-Daubechies-Feauveau (9,7) and Spline (5,3) biorthogonal filter bank, respectively [18].

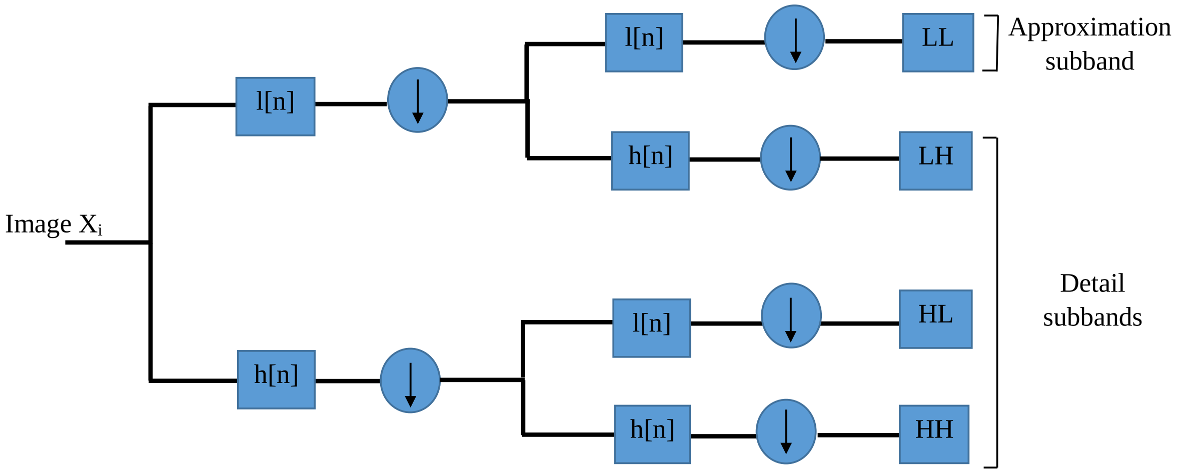


Figure 2.2: Block scheme of the low-pass and high-pass filter realization for an image X_i.

2.1.2 Quantization

Quantization step maps the wavelet coefficients values obtained after DWT to a smaller range of values [49]. This results in reduction of data precision and is used mainly with lossy image compression techniques. In JPEG 2000 compression standard, mainly two types of quantization techniques are available: i) scalar quantization, and ii) trellis coded quantization (TCQ) [18]. For lossy compression, each of the wavelet coefficients is quantized using a particular scalar value, while for the lossless compression the quantization step is neglected. For scalar quantization, the wavelet coefficient obtained after DWT is mapped to a smaller subset of values, depending on the step size used for the quantization that can be represented as:

$$q_{LL}(m,n) = sign(t_{LL}(m,n)) \left\lfloor \frac{t_{LL}(m,n)}{\Delta_{LL}} \right\rfloor \tag{2.3}$$

where $q_{LL}(m,n)$ represents the quantized value of the $t_{LL}(m,n)$ wavelet sub-band and Δ_{LL} represents the quantization step. In the case of lossless compression, the quantization step size is initialized as 1. Larger the quantization step size, higher the compression ratio.

Lossless compression preserves a good quality of the given input image whereas in the case of lossy compression higher compression ratio is achieved with degraded image quality. Before performing the entropy coding, each quantized wavelet sub-band is sub-divided into non-overlapping rectangular blocks called *precinct* and each *precinct* is further sub-divided into non-overlapping blocks called *code-blocks* that are represented as bit planes. Each *code-block* has usually size of 32×32 or 64×64 pixels. The size of code-blocks c is usually in powers of 2 and $c \leq 4096$. Fig. 2.1 shows the important blocks for JPEG 2000 compression algorithm. The bit rate control and resilience to the transmission error are the major benefits achieved through this code-block representation.

2.1.3 Entropy Block Coding with Optimized Truncation (EBCOT)

EBCOT, the entropy coding paradigm used in JPEG 2000 framework requires code-blocks of the wavelet sub-bands to generate the codestream of the given input data. EBCOT is subdivided into two steps: i) *Tier-1*; and ii) *Tier-2* encoding [18].

Tier-1 **Encoding**

In *Tier-1* encoding, each code-block associated with each wavelet sub-bands is entropy-coded using: i) Context Modelling; and ii) Arithmetic coding.

The *code-block* associated with each wavelet sub-band is represented in the form of bit-planes. Contextual information of bit planes of these code-blocks that is achieved by analyzing the neighborhood of these code-blocks can be obtained from three coding passes: *significance propagation pass*, *magnitude refinement pass* and *clean up pass.* In *significant propagation pass*, the bit is encoded if it is not significant or with at least one significant neighbor. In *magnitude refinement pass*, all the bits that are significant in the previous pass are coded and finally all the bits that are not coded are encoded in the *cleanup pass.* The contextual information of these code-blocks is encoded from Most Significant Bit (MSB) to Least Significant Bit (LSB) to obtain the compressed bit stream, which is performed in the *Tier-1* coding of EBCOT. The contextual information obtained from these coding passes are then encoded using arithmetic binary MQ-coder which results in a bitstream of a given input image.

Tier-2 **Encoding**

In *Tier-2* encoding, the compressed bit streams are organized into several *packets* and *layers* based on the resolution, component, spatial area and quality. *Packet* structure contains information about a few spatially consistent subgroups of *code-block* within a particular resolution, quality or level [18]. This packet structure organization allows to access the compressed bit stream of any resolution, level or component without decoding the entire compressed image. Each *packet* contains a header that provides information regarding the codestream associated with it. The packet structure provides information regarding the number of coding passes, zero-bit plane information, the size of the data obtained from a given code-block. This information is obtained in the *header* of a packet. One main header is associated with the compressed codestream of a given image. This main header includes information regarding the size of the image/tile, the number of wavelet decomposition used for each spectral band/image/tile, type of quantization used before entropy coding and several others. This freedom to access information regarding any level, resolution or component without decoding the entire compressed image is often termed as '*scalability*' . This arrangement allows a progressive encoding as well as decoding that can be utilized to address image retrieval or scene classification problems in JPEG 2000 compressed RS image archives.

2.1.4 Deep Neural Networks

Deep Neural Networks (DNNs) are a class of artificial neural networks (ANNs) that comprise of several non-linear operations that are used to learn hierarchical feature representations from a given set of images [24; 50].

A DNN generally comprises an *input layer*, *multiple hidden layers* and an *output layer*. A DNN architecture consists of a minimum of three types of layers that helped to add the term 'deep' in a DNN. At each hidden layer, a discriminative set of feature representations are learned based on the feature representations obtained from the preceding hidden layer. Feature representations obtained from higher hidden layers are more discriminative as they provide information such as edge, shape, etc. They have demonstrated their high feature learning capability in several domains such as visual recognition, object detection, change detection, classification and several others [28; 51]. These networks are inspired from the ANNs, a class of neural networks that were developed from the biological neural networks (which comprises human brain). Each layer constitutes a large number of *neurons* and the input data is fed to these neurons. The non-linearity to a DNN model is provided by the *activation functions* at each hidden layer. Each *neuron* in the hidden layer comprises of these *activation functions*, which are mathematical representations that determine the output of a particular neuron. Several types of activation functions were introduced to the literature such as Sigmoid, , Leaky ReLU, etc [52]. As an example, ReLU activation function can be defined as:

$$ReLU(z) = max\{0, z\} \tag{2.4}$$

where, z is the input to a neuron in the hidden layer. The data required for a given DNN model can be categorized as: i) *train*, ii) *validation* and iii) *test* data sets. The *training* set contains input data with their associated labels, which are used to learn the patterns of the data, which are validated using the *validation* set (where the parameters of the neural network is optimized). The *test* set contains a small subset of unlabelled data and performance of the neural network is finally evaluated on the test set. Learning of a given DNN can be categorised as: i) *supervised*; and ii) *unsupervised.* In a *supervised* learning scheme, the neural network is provided with labelled input data to train the model with learned representations and predicts the class labels of the test set. In an unsupervised learning scheme, the input data does not have an associated class label information and the model automatically learns the pattern of the data automatically.

The performance of a neural network model mainly depends on a *loss function*, which evaluates how well the given input dataset is modelled by the presented algorithm. The aim of the considered model is to minimize the loss function to efficiently reduce the distance between the training data and the expected outcome (class label). Mean squared error (MSE), cross entropy function are some of the most commonly used loss functions in the neural network models. The predicted outcomes obtained for all the input data do not alone efficiently model the considered neural network architecture. Thus, the generalisation capability of the proposed model is provided by the *regularization* function. *l1*, *l2*, *ridge* regularization are some of the most commonly used regularization methods. Although they have demonstrated impressive performance over several domains, still there is a need to address several challenges. The two main challenges that are faced by the use of DNNs are: i) overfitting; and ii) long computational time. Various DNN architectures such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) network were introduced to several RS domains. In the recent years, experiments have shown the feasibility of using pre-trained deep architectures for RS image scene classification tasks. Thus this approach gained increasing research interest for several RS applications.

2.1.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), are a class of DNNs that employ the 'convolution' operation in the hidden layers to perform a given task [24]. They are generally used when the input is an image and have demonstrated their capability in image feature learning and classification.

CNNs use a supervised learning scheme, where the network is trained using images with their associated class labels. Given a set of labelled input data, CNNs train the labelled data to learn the feature representations which are obtained using convolutional operations. The supervised learning scheme of CNN uses backpropagation algorithm to

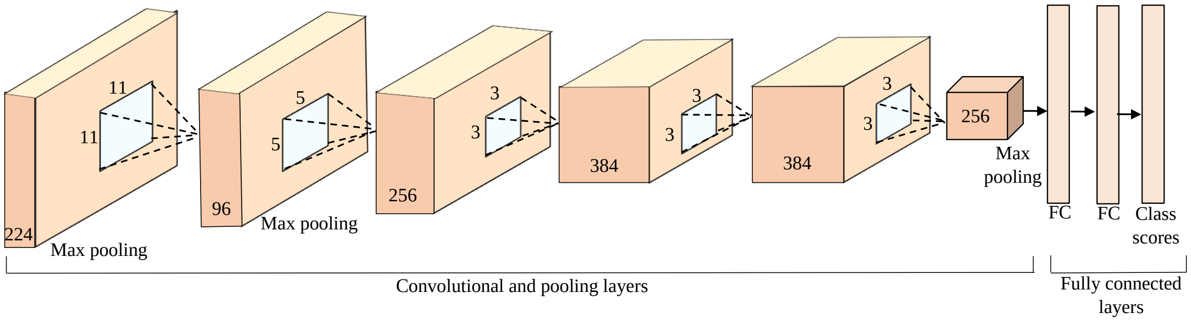


Figure 2.3: Block scheme of the AlexNet architecture.

minimize the error between the input and predicted class labels. Figure 2.3 demonstrates the block scheme of the AlexNet architecture. In the CNN architecture, one may notice convolution and pooling (hidden) layers and finally a Fully Connected (FC) layer. Each hidden layer in a CNN generally comprises of: i) convolutional; and ii) pooling layer.

Convolutional Layer

The convolutional layer obtains the feature representations of a given image that are obtained based on a convolution operation between the considered *filter* and the input image, which can be represented as,

$$Y_i = X_i \circledast K \tag{2.5}$$

where K represents the considered filter matrix, X_i represents the input image and Y_i represents the resultant matrix (*feature map*) obtained after the convolution operation. A *filter* can be considered as a weight matrix that is used to obtain representative features from the considered input image X_i. Each layer in a CNN represents varying features obtained from the image. During training, the weights are randomly initialized and then multiplied with the pixel values associated with the given filter. The filter of a given size slides around the considered image and obtains the feature maps associated with it. The pattern in which the filter slides across a given image is defined by the *stride*. If the value of *stride* is one, then the output of the convolution operation is downsampled by 1. In addition to the stride value, the convolution layer requires the *padding* value to the considered image. While performing convolution operation towards the edge of an image, if the boundaries of the filter k is outside the image, then the resultant matrix cannot be evaluated. To avoid this error, we add a zero-padding surrounding the considered image (rows and columns). *Padding* maintains the information at the edges of an image that is required during a convolution operation. The size of the resulting feature map is obtained as:

$$f_{size} = \frac{w - k + 2p}{s} + 1 \tag{2.6}$$

where f_{size} represents the size of the resultant feature map, w represents the width of the input data, k denotes the filter size, p and s denote the value of the padding and stride, respectively. CNNs are mostly considered with image recognition, classification tasks because the filter k is shared with many images at several positions which reduces the number of parameters required by the CNN network. The value of the filter usually varies from 2×2 to 9×9.

Pooling Layer

After the convolutional layer, the pooling layer results in the downsampled size of the feature maps obtained [24]. The downsampled feature maps represents the features obtained after the convolutional layer which are mostly invariant to translation. Generally, two types of pooling operations are used in CNNs: i) average pooling and ii) max pooling. Given a filter size, average pooling obtains the average of the pixel values in the considered filter of the feature map whereas max pooling obtains the maximum value of the feature map in the considered neighborhood. In addition to translation invariance, the pooling layer reduces the number of hyperparameters required which is achieved from the reduction in the spatial dimension of the feature maps obtained and also reduces the overfitting of the data.

Fully Connected (FC) Layer

The feature maps obtained after the convolution and pooling layers is represented in a 'flattened' form using a FC layer. If the aim is to perform scene classification of images associated with single labels, the values in these flattened vector represents the probabilities of the class labels associated with a given image. As an example, if the given image belongs to 'residential' category, the FC vector will have higher probability values associated with the residential class. The initial FC layers take as input the feature maps along with their corresponding weights to predict the associated class label. The final FC layer that consists of number of neurons equivalent to the number of classes in the considered dataset provides the estimated probabilities along with their class labels.

Chapter 3

State of the art in Compressed domain CBIR and Scene Classification

This chapter provides the state-of-the-art for CBIR in RS, existing approaches in JPEG 2000 compressed domain and scene classification of RS images using deep learning techniques.

In this chapter, we provide a state-of-the-art analysis of the existing image retrieval and scene classification approaches in RS. Existing studies to perform image retrieval and scene classification for real large-scale RS image archives requires full decompression of RS images which is a computationally demanding task to perform. The continuous growth in the amount of compressed RS images demands developing efficient methods to perform image retrieval and classification. Several operational satellites such as Sentinel 2, PRISMA uses JPEG 2000 algorithm that uses wavelet based compression technique to store their images. In computer vision and pattern recognition, few studies highlight the potential of developing novel retrieval and classification approaches in compressed domain which has not yet been addressed to RS images.

This thesis focuses on developing novel methods to achieve computationally efficient image retrieval and scene classification in compressed RS image archives. Thus, section 3.1 discusses the existing image retrieval techniques in both computer vision as well as RS. Section 3.2 provides the state-of-the-art methods for scene classification in RS using deep learning techniques. Section 3.3 discusses the existing methods that use JPEG 2000 compressed domain features in computer vision and pattern recognition.

3.1 Remote Sensing Content Based Image Retrieval Systems

In view of the limitations that exist in conventional RS image retrieval systems as well as the increasing availability of massive amount of compressed RS data, developing efficient CBIR systems in compressed domain is one among the major challenges faced by the RS society. A general RS CBIR system mainly comprises two steps: i) Feature Extraction; and ii) Similarity Assessment [21]. Performance of a given CBIR system mainly depends on:

1. effective modelling of *feature descriptors* for a given image, and
2. adept assessment of similarity between the query image and archive images.

Feature descriptors, which represents mathematical representations of a given image, can be mainly categorized into two i) *conventional* and ii) *learned*, based on the analysis of how they are obtained. Conventional (traditional) feature descriptors generally obtain spectral, texture, edge, shape or intensity information using a single (global) representation for a given image. Spectral histograms [53], one among the earliest and simplest global descriptor, is obtained by considering the marginal distributions obtained through the responses of a given filter applied to a given image. Although these descriptors are rotation and translation invariant, they are illumination sensitive [54]. Based on the advancing imaging technologies used to capture the RS data, images in the archive may be affected by varying illumination conditions and further leads to a poor performance of the CBIR system. Texture features, obtained using Grey level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP) and its variants, Gabor descriptor have been widely studied to address image retrieval problems in RS image archives. These representations are obtained from statistical analysis of a given image or local image regions. GLCM descriptor analyzes the inherent properties of a given image such as contrast, homogeneity, entropy, correlation coefficient, energy and several others [32; 55; 56; 57]. LBP and its several variants models patterns present in a given image by considering a pixel along with its surrounding neighborhood using binary codes has been found very effective in RS CBIR [58; 59; 60; 61]. Similar to LBP, Gabor filter characterizes a given image by analyzing the frequency distribution in the neighborhood of a given pixel [62]. Edge features such as Histogram of Gradients (HOG), mathematical morphological operators have also been studied to address image retrieval problems [36; 63; 64]. Global texture features obtained from varying image scales has contributed to an improvement in the performance when compared to the global image descriptors obtained at a single scale.

Considering the complexities in RS images, the performance of these aforementioned low-level global feature representations to address image retrieval problems in real large-scale archive is very low. In addition, these features do not consider the additional spatial or contextual information that exist within the sub-regions of high-resolution RS images which could hasten the performance of a given CBIR system. Thus, global representation of feature descriptors obtained from several local regions (local-global representations) were considered. One such representation is the Scale-Invariant Feature Transform (SIFT), which was proposed to capture the additional contextual information where the features are obtained from several key-points (interest-points) obtained over a given image [36; 65]. Further, their representation of bag-of-visual-words (BoVW) as well as extended BoVW were introduced to image retrieval problems in RS [36]. Graph based representations, where the nodes represent the region attributes and the edges represent the spatial relationship between the regions were also proposed in RS CBIR [21]. Hashing based methods were introduced where the images are represented using binary hash codes to reduce the memory requirements also received increasing attention to RS image retrieval problems [34; 66]. In addition to BoVW, Vector of Locally Aggregated Descriptors (VLAD) was also another locally aggregated global representation of the features obtained [67; 68].

In the recent years, the potential of DL methods to learn the high-level semantic content of RS images has shown remarkable improvements in the retrieval performance when compared to traditional descriptors for large-scale RS image archives. In [27], the retrieval system utilizes the features obtained from the CNN to perform image retrieval by considering the weight of each class in the given query image. Features obtained from a CNN network can be directly obtained from the feature maps obtained after convolution or max-pooling as well as from the FC layer. However, these high-dimensional features require huge storage requirements which incur additional costs. To address this, Zhou et al. [69] proposed a CNN networks that uses a three-layer perceptron to represent deep features at a lower dimension. DL methods were also presented in the context of hashing to generate semantically efficient CNN features and binary hash codes by considering cross-entropy loss [70]. To consider scale variation of images in the archive, multi-scale CNN (MCNN) was proposed to model RS images where representational deep features are obtained [71]. However, training each CNN separately for images at varying scales is cost-ineffective and computationally-demanding.

After obtaining the feature descriptors, the next step is to assess the similarity between the query image and the archive images. One of the simplest measures used to calculate the similarity is the k-nearest neighbour (k-nn) approach where the feature descriptor of a given query image is used to obtain the first k best matches from the image archive [22]. The similarity can also be computed using the distance measures such as Euclidean distance, cosine similarity, Minkowski, Histogram Intersection (HI), Kullback Leibler Divergence (KLD) and several others. When the feature descriptors obtained are histogram representations, HI is commonly used. For statistical representations obtained from the images KLD, earth movers distance, wasserstein distance and several others measures are used [72]. When RS images are modelled using graphs, methods such as inexact graph matching strategy as presented in [21] is used to assess similarity. Image retrieval can also be considered as a binary classification problem where there is a subset of relevant and irrelevant images to a given query image. In this case, we can consider image retrieval as a binary classification problem. Binary classifiers such as Support Vector Machines (SVMs), Maximum Likelihood Classifiers (MLCs) were used to to address RS image retrieval [73; 74]. When RS images are represented using binary hash codes, hamming distance measure is used to calculate the similarity [75; 76]. Spatial pyramid matching (SPM) technique was proposed to model the local information of the images obtained at several resolutions [77]. The local information obtained using BOVW histogram representation from several resolutions are concatenated together to form the SPM feature descriptor. However, they do not model the spectral information within the RS images and thus, spatial-spectral pyramid matching (SSPM) technique was proposed to classify hyperspectral RS images [78]. To reduce the classifiers error which are used to improve the retrieval performance, Relevance Feedback (RF) was proposed where the user provides feedback to examine the retrieved results [33].

3.2 Scene Classification using DL approaches in RS

In the recent years, the potential of DL methods to perform scene classification of RS images gained huge popularity due to its ability to learn the underlying high-level semantic content of the images. Several efforts were carried out to develop effective scene classification approaches in RS image archives. The main goal of any scene classification task is to assign a class label to an image based on the analysis of the feature representations obtained from them. Scene classification for RS images was conventionally performed using several handcrafted features such as LBP, spectral histogram, GLCM and several others. However, the performance of these handcrafted features are very shallow and require human intelligence to obtain them which is a computationally demanding task as well as incur additional labor costs. Recently, DL approaches has shown remarkable improvements in performance in several domains in particular image scene classification.

CNN is one among the most popular DL algorithms that has shown its ability to learn high-level semantic patterns in RS images and has demonstrated impressive improvements in performance over traditional scene classification approaches. In the early years, training the CNN model from scratch was considered to perform RS image scene classification. However, recently, it is shown that the use of pretrained models such as AlexNet [79], GoogleNet [80], VGG16 [81], CaffeNet [82] for RS images has helped to improve the scene classification performance [37]. Two limitations arise while using pretrained models in CNNs: i) image size constraint; and ii) overfitting problem, which leads to a decrease in classification performance. In addition, the features obtained from high-level convolutional layers are abstract and may not efficiently characterise objects present in the considered image. To address this, Guoli et al. [83] proposed a novel approach that integrate the deep features obtained from intermediate convolutional layers to improve the discriminative power of the feature descriptors and classification accuracy. However, this approach do not consider the images at varying resolutions which are present in large-scale RS image archives. In view of this, Zheng et al. [26] proposed a deep representation where the feature representations at multiple scales are obtained from the image feature maps using multiscale pooling (MSP) to improve the classification performance. Also, several efforts were also made to obtain features obtained at varying resolutions using parallel CNNs [84]. However, the training time required for these parallel CNNs are computationally demanding and is less efficient when compared to the standard CNNs. To reduce the number of parameters included in the FC layer, Boualleg et al. [85] introduced the novel approach where the parameters obtained from the FC layer is reduced using deep forest classifier. In CNN model, one may consider that the features obtained from the final convolutional layer and FC layer depicts the local and global information obtained from the image. In [86], the proposed approach considers the combination of global and rearranged local features to obtain representation with higher discriminative power to perform scene classification.

Another notable class of DNN model are the Autoencoders (AE) that learns the compressed image representations to perform scene classification [23; 40; 43]. Chen et al. [25] proposed a novel DL architecture that uses stacked AEs to obtain highly representative descriptors by combining the spatial information and deep features for hyperspectral images (HSI). They have shown remarkable improvement in classification performance over traditional HSI classification methods. Following this, Ma e al. [87] proposed a framework where both spatial and spectral information obtained from the HSI are considered for classification. In [88], a two stream DNN framework was proposed where spectral and spatial information are fed separately to perform HSI classification. They proposed a weighted class probability fusion scheme to assign weights to features obtained from two separate streams and has shown improvement when compared to state-of-the-art methods. However, considering the time demanding task of considering many hypercubes in the nodes of hidden layers, a segmented stacked AE was proposed to handle subset of hypercubes at each segments [89]. To perform unsupervised deep feature extraction, sparse AEs was considered. In [39], a hierarchical convolutional sparse AE was proposed that takes as input image patches to achieve discriminative features by considering the feature maps obtained after pooling. Due to the limited availability of labeled samples, a novel architecture that considers a siamese network and AE was proposed to perform scene classification [90]. They utilised the ability of siamese network to increase the number of training samples was utilized and AEs to learn the high-level semantic content. DL models require large number of annotated training samples to learn the semantic structure of the data. Training models with limited number of training samples may result in overfitting of the data. Inorder to alleviate the problem of overfitting that occur due to the limited availability of training samples, recently Generative Adversarial Networks (GANs) was introduced to RS scene classification problems.

In the recent years, GANs have demonstrated massive success in several RS domain in particular scene classification[29; 42; 91; 92]. They generally have a generator (that learns and generates the semantic content of the input data) and a discriminator (that classifies the generated as well as the input data) network. MARTA GAN [93], was one among the initial efforts made to exploit the performance of GAN in RS domain. They learn the mid level features and global feature matching approach to improve the performance of the state-of-the-art DL approaches. Zhu et al. [94] proposed a 1-D and 3-D GANs to classify spatial and spatial-spectral information of hyperspectral images, respectively. They use combination of two CNNs to generate and classify the hyperspectral images. The performance of a given GAN model mainly depends on its ability to efficiently generate the images from the limited amount of given input data. In [30], an NL-GAN was proposed to incorporate the non-local spatial information of the images to train the GAN model and has shown improvement in the classification performance. In [95], a semi-supervised hyperspectral GAN was proposed to achieve encouraging classification performance with only limited number of labelled samples. The performance of GAN model mainly depends on its ability to generate images as well as its discriminative capability in the generator and discriminator network, respectively. To address these issues, a multiclass spatial-spectral GAN (MSGAN) was proposed in [96]. MSGAN also considers the spatial-spectral information that is ignored otherwise in GAN models. The potential of CNNs to learn high-level semantic representations of the image samples has led to the introduction of deep convolutional GANs (DCGAN) [97]. Although GANs has proved its ability to perform efficient scene classification, the training as well as optimization of generator and discriminator network is computationally demanding when compared to other DL models.

Recurrent Neural Networks (RNNs) is another branch of DL models that has become popular in RS domain due to its ability to remember and predict past and future instances, respectively [98; 99; 100; 101]. It has been widely used in RS time series studies and has recently become popular to address scene classification problems. In Mou et al. [31], the proposed approach considers RNN model for the first time to perform scene classification for hyperspectral images. They propose a novel parametric named as *rectified tanh* activation function to analyze sequential hyperspectral data and has achieved remarkable classification performance. Although RNN has demonstrated good scene classification performance, they endure gradient vanishing problem that deteriorated its capability to learn the previous instances. To address this, Long Short-Term Memory (LSTM) network was introduced to improve the overall performance. To further explore the co-occurrence relationship between various classes, a bidirectional attention network was proposed in [102]. In [41], a novel framework is considered where the high-level semantic as well as spatial features are given attention to improve the classification accuracy. In the considered work, they propose a method to reduce the required number of parameters of the DL model by considering simple vector of the features obtained to feed into a network that considered relevant image regions required for classification. The fact that the classification performance could be improved by considering the relevant areas within an image has encouraged to introduce attention mechanisms. Attention mechanism learn discriminative relevant image features from specific region by avoiding the redundant information obtained from irrelevant image regions. Sumbul and Demir [103] proposed a novel attention scheme that considers the local image descriptors obtained from relevant image regions. The fact that the classification performance could be improved by considering local as well as global features is utilized in a novel local-global attention framework is considered in [38]. Despite the computational complexity endured in training the DL models, they have shown impressive classification performance and has huge potential to address many research problems.

3.3 Existing JPEG 2000 based feature descriptors

Although all the aforementioned scene classification and retrieval methods (Section 3.1 and 3.3) has shown remarkable performance, they require full decompression of the images before performing the retrieval or classification tasks. In computer vision and pattern recognition, several notable efforts were carried out to model the partially-decoded compressed image representations obtained while using JPEG 2000 algorithm. When JPEG 2000 is considered, two types of feature descriptors can be obtained: i) header-based; and ii) wavelet-based features [104]. Header-based features are obtained directly from the codestreams of a given image while wavelet-based features are obtained from the partially-decoded wavelet coefficients (which are obtained from the sub-bands). Zargari et al. [105] proposed an image retrieval approach where feature descriptors such as Maximum Bit Histogram (MBH), Importance Histogram, Compression Rate Vector are obtained directly from the packet header information associated with a given code-block. In addition, it is also possible to obtain the number of entropy coded bytes that were used to encode a given code-block associated with a particular wavelet sub-band. In [106], they use the number of zero bitplanes present in a given a code-block. However, the discrimination power associated with these features are not effective when compared to the features that are obtained from the partially decoded wavelet coefficients. In the past decade, many efforts were put-forth to obtain efficient features from the wavelet sub-bands and has shown to be effective to perform image classification/retrieval tasks.

The simplest and the most widely used texture descriptor that can be obtained from the wavelet sub-band is the energy and mean descriptor which is obtained by calculating the sum of squares and obtaining the mean of all the wavelet coefficients. However, energy descriptor neglects the semantic contents of the RS image and results in a very low retrieval performance. In some studies, attempts were carried out to obtain the HOG, GLCM, LBP descriptors from each wavelet sub-band for face recognition as well as classification tasks [107; 108; 109]. In [110], the proposed approach obtains histogram of the local energy calculated at several patches of the wavelet sub-bands to perform texture classification. However, calculating these descriptors from all the wavelet sub-bands at each resolution and finally aggregating them is time-demanding as well as less efficient when considering real large-scale RS image archives. Several efforts were made to use mathematical morphological operations such as dilation and erosion when applied to images allows to obtain the shape information [56; 63]. Later, it was observed that all the detail wavelet sub-bands demonstrate a near-Gaussian behavior and thereby several efforts were put-forth to model them using Generalized Gaussian Distribution (GGD) [111; 112]. Teynor et al. [53] attempts to model the detail wavelet sub-bands using Gaussian Mixture Model (GMM) and approximation coefficients using the color histograms. They observed that the performance of GMM is superior over GGD for CBIR. Another variant of GGD and GMM was Generalized Gamma Distribution (GΓD) used for image classification and retrieval [113]. Although these statistical representations works efficiently, they are computationally expensive and time demanding. They combine the edge information obtained from the moduli of the horizontal and vertical wavelet coefficients with the angle or the orientation of the images that is obtained using the tangent of the wavelet coefficients of the horizontal and vertical wavelet sub-bands at each resolution. Considering the scenario, there are no related efforts that were made to address the classification or retrieval of images from JPEG 2000 compressed RS image archives.

Chapter 4

A Progressive Content Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives

In this chapter we present a novel CBIR system that achieves a coarse to fine progressive RS image description and retrieval in the partially decoded JPEG 2000 compressed domain. The proposed system initially: i) decodes the code-blocks associated only to the coarse wavelet resolution, and ii) discards the most irrelevant images to the query image based on the similarities computed on the coarse resolution wavelet features of the query and archive images. Then, the code-blocks associated to the sub-sequent resolution of the remaining images are decoded and the most irrelevant images are discarded by computing similarities considering the image features associated to both resolutions. This is achieved by using the pyramid match kernel similarity measure that assigns higher weights to the features associated to the finer wavelet resolution than to those related to the coarse wavelet resolution. These processes are iterated until the codestreams associated to the highest wavelet resolution are decoded. Then, the final retrieval is performed on a very small set of completely decoded images. Experimental results obtained on two benchmark archives of aerial images point out that the proposed system is much faster while providing a similar retrieval accuracy than the standard CBIR systems.

Part of the chapter appears in:

1. Byju, A. P., Demir, B., Bruzzone., ’ A Progressive Content Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives’ , IEEE Transactions on Geoscience and Remote Sensing, accepted for publication, 2020.

4.1 Introduction

Recent developments in satellite technologies witnessed a massive accumulation of huge amounts of data (petabytes) in RS image archives. Effective utilization (such as storage, management and retrieval) of such huge amounts of data has become one among the challenging issues faced by the RS community. Developing efficient solutions to effectively exploit these data using CBIR techniques is one of the most researched topics in RS. However, in order to reduce the necessary storage requirements, RS images are compressed before being stored in any archive [7; 8; 9; 10; 11; 12; 13]. This intensifies the challenges involved in retrieving images from large-scale compressed RS image archives. In computer vision and pattern recognition, few efforts were made to develop image retrieval techniques in compressed image archives. Although most of the existing RS CBIR systems work efficiently (see Section 3.1), they require fully decompressed images as input to the system to perform the image retrieval task. Considering the gigabytes worthy images that are stored per day, applying decompression, obtaining feature descriptors and assessing similarity to each image in the archive is impractical in real large-scale RS image archives. In view of this, it is crucial to develop efficient image retrieval techniques that: (i) minimize the amount of decompression required for the images; and (ii) achieve similar performance when compared to the CBIR systems that require fully decompressed images.

In this chapter, we present a novel progressive CBIR system that performs image retrieval in compressed RS image archives. We assume that the images in the archive are compressed using the JPEG 2000 compression algorithm. In view of this, the proposed system aims to minimize the amount of decompression applied to the images before the retrieval is performed in real large-scale RS image archives. Here, we present a novel system that: i) initially takes the codestreams associated with the coarsest wavelet resolution to obtain the feature descriptors; ii) performs a weighted kernel based similarity assessment using pyramid match (PM) kernel to discard irrelevant images using the set of features obtained. Further, code-blocks at a finer level associated with the remaining subset of relevant images are estimated to compute the similarity considering the descriptors obtained from both previous as well as the current wavelet resolution to identify the first relevant images to be retrieved. The proposed system is completely unsupervised and can be adapted to any image descriptor that can accurately describe wavelet coefficients. Experimental results obtained on two benchmark archives demonstrate the effectiveness of the proposed system.

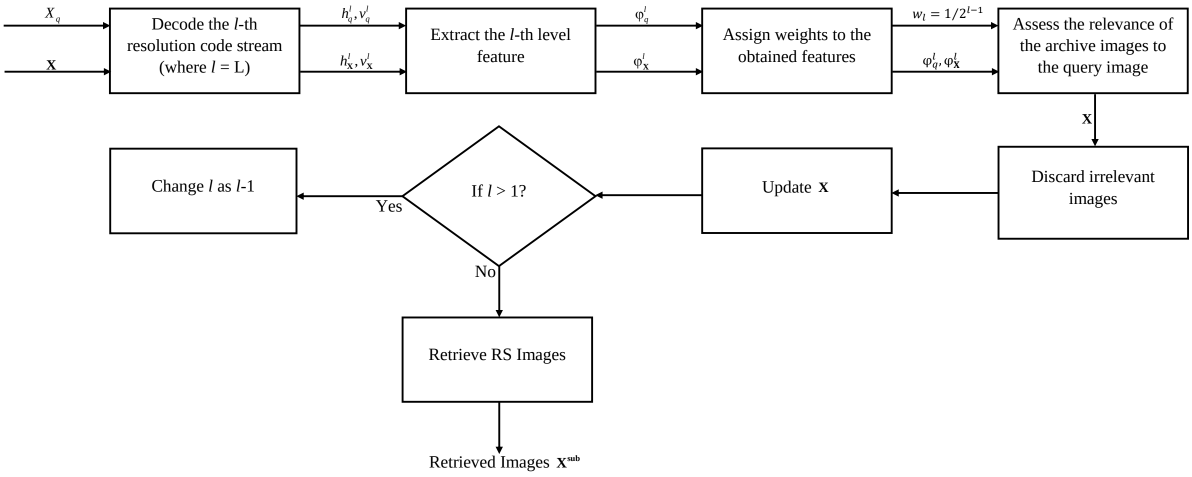

X_q
$\mathbf{X}$
Decode the l-th resolution code stream (where l = L)
h_q^l, v_q^l
$h_{\mathbf{X}}^l, v_{\mathbf{X}}^l$
Extract the l-th level feature
φ_q^l
$\varphi_{\mathbf{X}}^l$
Assign weights to the obtained features
$w_l = 1/2^{l-1}$
$\varphi_q^l, \varphi_{\mathbf{X}}^l$
Assess the relevance of the archive images to the query image
$\mathbf{X}$
Discard irrelevant images
Update $\mathbf{X}$
If l > 1?
Yes
Change l as l-1
No
Retrieve RS Images
Retrieved Images $\mathbf{X}^{sub}$

Figure 4.1: Block scheme of the proposed coarse to fine progressive RS CBIR system within the JPEG 2000 framework.

4.2 Proposed CBIR system

4.2.1 Problem formulation

Let $\mathbf{X} = \{X_i\}_{i=1}^N$ be an archive that consists of a large number of N JPEG 2000 compressed RS images, where X_i represents the i-th compressed image. Given a query image ($X_q \in \mathbf{X}$ or $X_q \notin \mathbf{X}$) the main objective of the proposed system is to retrieve a set of relevant images $X^{rel} \subset \mathbf{X}$ from the archive $\mathbf{X}$ that are semantically similar to X_q without fully decoding all the images in $\mathbf{X}$. We assume that images in the archive are compressed by using the JPEG 2000 algorithm based on L wavelet decomposition levels, i.e. an image X_i has one low-pass sub-band (approximation sub-band) and *3L* high-pass (horizontal, vertical and diagonal) sub-bands. Thus, the total number of sub-bands for a given image with L decomposition levels is $3L$+1. When JPEG 2000 is considered, the simplest approach to perform image retrieval consists of three main steps: 1) entropy-decoding of all the images in the archive $\mathbf{X}$, 2) extraction of image descriptors, and 3) analysis of similarity and retrieval of images relevant to the query image. However, entropy-decoding all the N images up to L decomposition levels in a large-scale image archive is time-consuming and computationally challenging. To address this problem, we present a novel CBIR system that achieves a coarse to fine progressive RS image description and retrieval in the partially decoded JPEG 2000 compressed domain. Fig. 4.1 shows the general block scheme of the proposed system. In the following sub-sections, we initially introduce the method used for characterization of wavelet decomposition levels and then provide detailed explanation on the proposed CBIR system.

4.2.2 Characterization of Wavelet Decomposition Levels

In this chapter, we characterize each wavelet decomposition level with the texture descriptors proposed in [114] as the magnitude of wavelet frame coefficients. Let $h^l_{X_i}$ and $v^l_{X_i}$ be the horizontal and vertical sub-bands of an image X_i at the l-th wavelet decomposition level. To define the texture descriptor $H^l_{X_i}$ of the l-th level, the moduli $\varphi^l_{X_i}(u,v)$ of the horizontal and vertical detail coefficients are initially calculated as follows:

$$\begin{aligned}\varphi^l_{X_i}(i,j) = \sqrt{[h^l_{X_i}(i,j) + v^l_{X_i}(i,j)]^2}, i = 1, 2, ...m;\\ j = 1, 2, ...n\end{aligned} \tag{4.1}$$

where $h^l_{X_i}(i,j)$ and $v^l_{X_i}(i,j)$ represent horizontal and vertical coefficients, respectively, which are associated to the sample location (i,j) for the l-th sub-band of $m \times n$ size. Then the histogram $H^l_{X_i}$ of $\varphi^l_{X_i}(i,j), i = 1, 2, ...m; j = 1, 2, ...n$, which models the distribution of the moduli obtained from the sum of squares of horizontal and vertical wavelet sub-band coefficients, is taken as the descriptor of the l-th wavelet resolution. In order to estimate the histogram $H^l_{X_i}$, initially the range of possible values are defined by the minimum and maximum sample values of $\varphi^l_{X_i}$ (independently from the other wavelet decomposition levels) and the range is divided into r histogram intervals (i.e., r histogram bins). Then, the histogram $H^l_{X_i}$ is computed by counting how many of the patterns belong to each interval. Accordingly, the histogram describes a feature vector consisting of a marginal distribution of moduli of wavelet horizontal and vertical detail coefficients. Note that if a sufficient number r of histogram bins is defined, the histogram can represent the underlying distribution with a high precision. Thus, the histogram associated with each wavelet decomposition level of each image is capable of effectively capturing the texture content of the related image. It is worth noting that texture descriptors obtained from the lowest wavelet resolution are able to capture the global structure (coarse-scale objects) of an image, whereas the texture descriptors obtained from the higher wavelet resolutions are able to capture local detailed information (fine-scale objects). It is worth noting that the texture descriptor is a histogram-based descriptor and thus rotation and translation invariant. However, it is not scale invariant and does not explicitly model the different illumination conditions. In any case, the proposed system is independent from the selected descriptor and any descriptor that can accurately describe the wavelet coefficients can be used.

4.2.3 Proposed Progressive CBIR System

The proposed progressive CBIR system initially decodes the code streams associated with the lowest wavelet resolution (i.e., L-th level) for all N images in the archive and then extracts the texture descriptor $H^L_{X_i}$ [where l=L in (4.1)] that models the marginal distribution of moduli of wavelet horizontal and vertical detail coefficients at the L-the level. Then, the similarities between the descriptors $H^L_{X_i}$ and that of the query image $H^L_{X_q}$ are measured by the Histogram Intersection (HI) kernel that is defined as [115]:

$$HI(H^L_{X_q}, H^L_{X_i}) = \sum_{i=1}^{r} min(H^r_{X_q}, H^r_{X_i}) \tag{4.2}$$

where r represents the number of histogram intervals. Then, the most dissimilar images to the query image, which are associated to the lowest similarity values, are discarded and $\mathbf{X}$ is updated. The next step starts by: i) decoding the code-blocks associated to the subsequent resolution level (i.e., l=L-1) of the remaining images in the archive and the query image; and ii) extracting their texture descriptor $H^{L-1}_{X_i}$. Accordingly, each remaining image in $\mathbf{X}$ and the query image are represented by increasingly fine descriptors associated to the first two wavelet resolutions. It is worth noting that descriptors associated to higher wavelet resolutions are capable of modeling more detailed information of the fine-scale objects in the images with respect to those associated to the lower wavelet resolutions. Thus, while estimating similarities between the query image and remaining images, we give higher weight values w_{L-1} to the descriptors associated to the (L-1)-th wavelet resolution than to weight values w_L of the descriptors associated to the coarsest L-th wavelet resolution. This is done by using the pyramid match (PM) kernel similarity measure [116], which computes the weighted sum of all the implicit correspondences between the texture descriptors of the different wavelet decomposition levels by both considering their weights and preserving their individual distinctness at each level. The PM kernel takes a weighted sum of the number of matches (i.e., the number of samples that fall into the same histogram interval) that occur at each level of resolution, by assigning higher weights to the matches found at higher resolution with respect to those found at coarser resolutions. The PM kernel is defined as [116]:

$$PM(H^l_{X_q}, H^l_{X_i}) = \sum_{m=l}^{L} w_m N_m, with\, l < L \tag{4.3}$$

where, as suggested in [116], $w_m = 1/2^{m-1}$ and N_m shows the implicit partial correspondence between any two successive wavelet decomposition levels. Note that the number of matches found at level L can also include all the matches found at the finer level (L-1). Thus, the number of new matches found at level L is given by:

$$N_m = HI(H^m_{X_q}, H^m_{X_i}) - HI(H^{m-1}_{X_q}, H^{m-1}_{X_i}) \tag{4.4}$$

where m = 1,2,...L denotes the wavelet decomposition level. It is worth noting that the PM kernel similarity measure is presented in [116] to assess the implicit partial matching correspondences between two multiresolution histograms to achieve a discriminative classification of variable feature sets. In this chapter, we exploit it for estimating the similarity among the image descriptors that are associated to different wavelet decomposition levels.

After estimating the PM similarities, the most irrelevant images are discarded and **X** is updated. Then next step starts by decoding the code-blocks associated to the subsequent resolution (i.e., l=L-2) of the remaining images in **X** and describing the images by increasingly fine descriptors associated to the three wavelet resolutions. The image similarities are estimated by (4.4) including the three descriptors with their associated weight values, and most irrelevant images are discarded. These decoding and discarding processes are iterated until the code streams associated to the highest wavelet resolution (i.e., when l=1) are decoded. Then the most similar images to the query are selected. If the images in the archive are decomposed up to L wavelet levels, then the number of stages that discards irrelevant images in the proposed system will be L-1. Due to the progressive coarse to fine CBIR mechanism, the proposed system exploits a multiresolution and hierarchical feature space to accomplish a progressive RS CBIR with an optimal use of resources in terms of retrieval and decoding time. It is also worth noting that in the final retrieval of the proposed CBIR system using the fine features, any search strategy can be adopted.

4.3 Dataset description and experimental setup

To evaluate the effectiveness of the proposed system, we performed several experiments on two benchmark archives. The first one is the widely used UCMERCED benchmark archive that consists of 2100 images of size 256 × 256 pixels selected from aerial orthoimagery with a spatial resolution of 30 cm [36]. Images are obtained from USGS National Map Urban Area Imagery collection of the following U.S. regions: Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson and Ventura. To evaluate the performance of the proposed method, we considered the annotations of the images with multi-labels. The total number of the multi-labels is 17 (which are: airplane; bare-soil; buildings; cars; chaparral; court; dock; field; grass; mobile-home; pavement; sand; sea; ship; tanks; trees; water), while the number of labels associated with each image varies between 1 and 7 [117]. For the example of images with their associated multi-labels the reader is referred to [117].

The second archive is the AID benchmark archive that consists of 10,000 aerial images

Annotations are available at ‘http://bigearth.eu/datasets.html’ .

of size 600×600 pixels with spatial resolution variable between 0.5 m. and 8 m. To assess the effectiveness of the proposed system, we considered the annotations of the images with single labels. The total number of single labels is 30 (i.e., airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct). For examples of images and their labels the reader is referred to [37].

To assess the effectiveness of the proposed system, the images of both archives were initially compressed by the JPEG 2000 algorithm by using 3 wavelet decomposition levels (i.e., $L = 3$). It is worth noting that since the size of the code-block used to obtain the JPEG 2000 compressed codestream must not be less than 32×32 pixels to obtain relevant information from the compressed images, in both archives it is not possible to use $L > 3$. Each sub-band is represented by a 24-dimensional feature descriptor. After decoding the code streams associated to the lowest decomposition level (l=3), T_1 of the most irrelevant images are discarded, where T_1 represents the percentage of discarded images. Then, T_2 of the most irrelevant images are discarded after decoding the second lowest decomposition level (l=2), where T_2 represents the percentage of discarded images at the second level. Finally, the image retrieval is performed based on the k-nearest neighbor (k-nn) search strategy by using jointly with the features obtained from the highest wavelet decomposition level and the previous levels from the remaining subset of relevant images.

Results of each system for the UCMERCED archive are provided in terms of: i) average recall, ii) average precision, and iii) average computational time obtained in 2100 trials performed with 2100 selected query images from the archive. For the details on how the recall and precision are calculated in the framework of multi-label image search and retrieval problems, the reader is referred to [117]. The results obtained for the AID archive are provided in terms of (i) average precision, and (ii) average computational time associated to 10,000 trials with 10,000 selected query images from the archive. Note that while we estimate the average precision and recall for multi-label image retrieval, for the single label case, average precision and recall reduce to the same performance measures as that of the multi-label image retrieval. Thus, we report only the average precision values for the single label retrieval experiments. The retrieval performance for both archives was assessed on the top-20 retrieved images. All the experiments are implemented via MATLAB® on a standard PC with Intel®Xeon®CPU i3-6100 @ 3.40GHz, 16GB RAM.

4.4 Experimental Results

We carried out several experiments in order to: 1) compare the effectiveness of the considered descriptor that models the distribution of moduli of the horizontal and vertical detail coefficients (called as the DMHV descriptor hereafter) with respect to the popular descriptors that model the wavelet coefficients; 2) performance analysis with respect to varying values of T_1 and T_2 of discarded images after decoding code streams associated with the first two wavelet decomposition levels; and 3) evaluate and compare the effectiveness of the proposed system with respect to (i) a standard-CBIR system using SIFT features obtained from fully-decoded images; (ii) a standard-CBIR system using DMHV descriptors without coarse to fine strategy.

4.4.1 Comparison of the image descriptors in the compressed domain

In the first set of trials, we analyze and compare the effectiveness of the DMHV descriptor with the widely used descriptors adapted with wavelet coefficients in the literature. The selected descriptors are: 1) the extended energy signature (EES) [118]; 2) the local binary pattern (LBP) [59]; 3) the gray level co-occurrence matrix (GLCM) based measure [57]; 4) the joint use of the EES and the LBP; and 5) the local energy histogram (LEH) [110]. To have a fair comparison, we applied the DMHV descriptor to the entropy decoding of all wavelet decomposition levels (i.e., the coarse to fine retrieval strategy is not considered). Tables 4.1 and 4.2 show the results obtained for the UCMERCED and AID archives, respectively. From the tables, one can observe that the DMHV descriptor provides the highest accuracy for both archives. This is achieved at the cost of increasing the required computational time. The GLCM-based descriptor provides the second best performance. In detail, the DMHV descriptor results in an improvement of almost 6.34% and 6.86% in average precision and average recall, respectively for the UCMERCED archive when compared to the GLCM based descriptor with slightly higher computational time.

Fig. 4.2 and 4.3 show an example of images retrieved from the UCMERCED and AID archives, respectively, by considering all the above-mentioned descriptors. In Fig. 4.2 the query image includes *bare soil*, *buildings*, *cars*, *pavement* and *trees*. The retrieval order and the multi-labels associated with each image are given above and below the related image, respectively. By analyzing the figure one can observe that all the images retrieved by using the proposed DMHV descriptor [see Fig. 4.2(g)] contain almost all the class labels included in the query image. On the contrary, the images retrieved by using the other descriptors mostly contain only one or two of the class labels [see Fig. 4.2(b-f)]. In Fig. 4.3 the selected query image is from *dense residential* category and the retrieved images with their associated single labels are provided below the related image. By analyzing

Table 4.1: Comparison of the performance for different descriptors (UCMERCED archive).

Descriptors	Average Precision (%)	Average Recall (%)	Feature Extraction time (seconds)
EES [18]	59.80	61.73	7.11
LBP [59]	47.76	47.79	9.08
GLCM [57]	61.84	64.01	34.28
EES and LBP	59.24	60.84	20.97
LEH [110]	59.80	62.18	11.07
DMHV	**68.18**	**70.87**	**29.08**

Table 4.2: Comparison of the performance for different descriptors (AID archive).

Performance Metric	EES [18]	LBP [59]	GLCM [57]	EES and LBP	LEH [110]	**DMHV**
Average Precision (%)	51.34	49.97	52.97	51.17	50.57	**59.97**
Feature Extraction Time (seconds)	23.14	30.61	73.91	65.73	40.57	**75.59**

Table 4.3: Average precision and recall for the proposed progressive coarse to fine RS CBIR system at each level when T_1=25% (UCMERCED archive).

Decomposition Levels	Average Precision(%)	Average Recall(%)
Level 1 **(Coarsest Feature)**	65.04	67.09
Level 2 (Fine Feature)	67.76	67.79
Level 3 **(Finest Feature)**	68.28	70.94

the figure one can observe that all the images retrieved by using the DMHV descriptor [see Fig. 4.3(g)] belong to the *dense residential* category. When the other descriptors are used, some irrelevant images associated with category labels *airport*, *baseball field* and *sparse residential* are retrieved. By a visual analysis of all these results, we observe that the DMHV descriptor accurately models the content associated with each query image, resulting in retrieval of the visually most similar images from the archive.

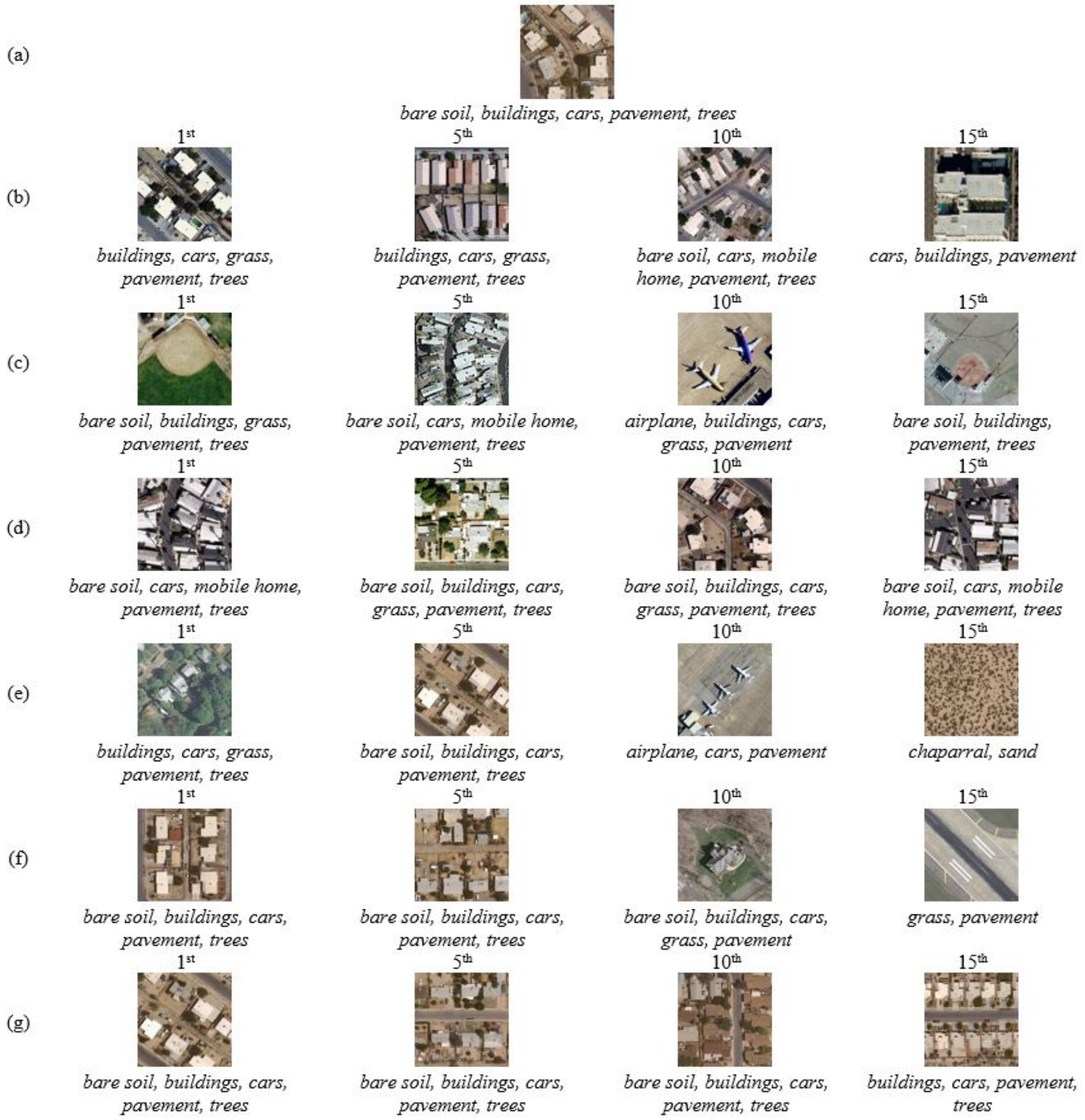

(a)
bare soil, buildings, cars, pavement, trees
1st
5th
10th
15th
(b)
buildings, cars, grass, pavement, trees
buildings, cars, grass, pavement, trees
bare soil, cars, mobile home, pavement, trees
cars, buildings, pavement
(c)
bare soil, buildings, grass, pavement, trees
bare soil, cars, mobile home, pavement, trees
airplane, buildings, cars, grass, pavement
bare soil, buildings, pavement, trees
(d)
bare soil, cars, mobile home, pavement, trees
bare soil, buildings, cars, grass, pavement, trees
bare soil, buildings, cars, grass, pavement, trees
bare soil, cars, mobile home, pavement, trees
(e)
buildings, cars, grass, pavement, trees
bare soil, buildings, cars, pavement, trees
airplane, cars, pavement
chaparral, sand
(f)
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, grass, pavement
grass, pavement
(g)
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, pavement, trees
buildings, cars, pavement, trees

Figure 4.2: Example of (a) query image; and retrieved images by using (b) the EES descriptor, (c) the LBP descriptor, (d) the GLCM descriptor, (e) a combination of the EES and the LBP descriptors, (f) the LEH descriptor and (g) the DMHV descriptor (UCMERCED archive).

4.4.2 Performance of the proposed system versus T_1 and T_2 values

In this subsection, we analyze the performance of the proposed progressive-CBIR system with respect to the parameters T_1 and T_2. Tables 4.3 and 4.4 report the performance measures obtained after performing pyramid match kernel similarity measure using features obtained from each wavelet decomposition level for UCMERCED and AID archives respectively. By analyzing the tables, one can see that there is a significant improvement in the performance measures when hierarchical weights are assigned to the progressively

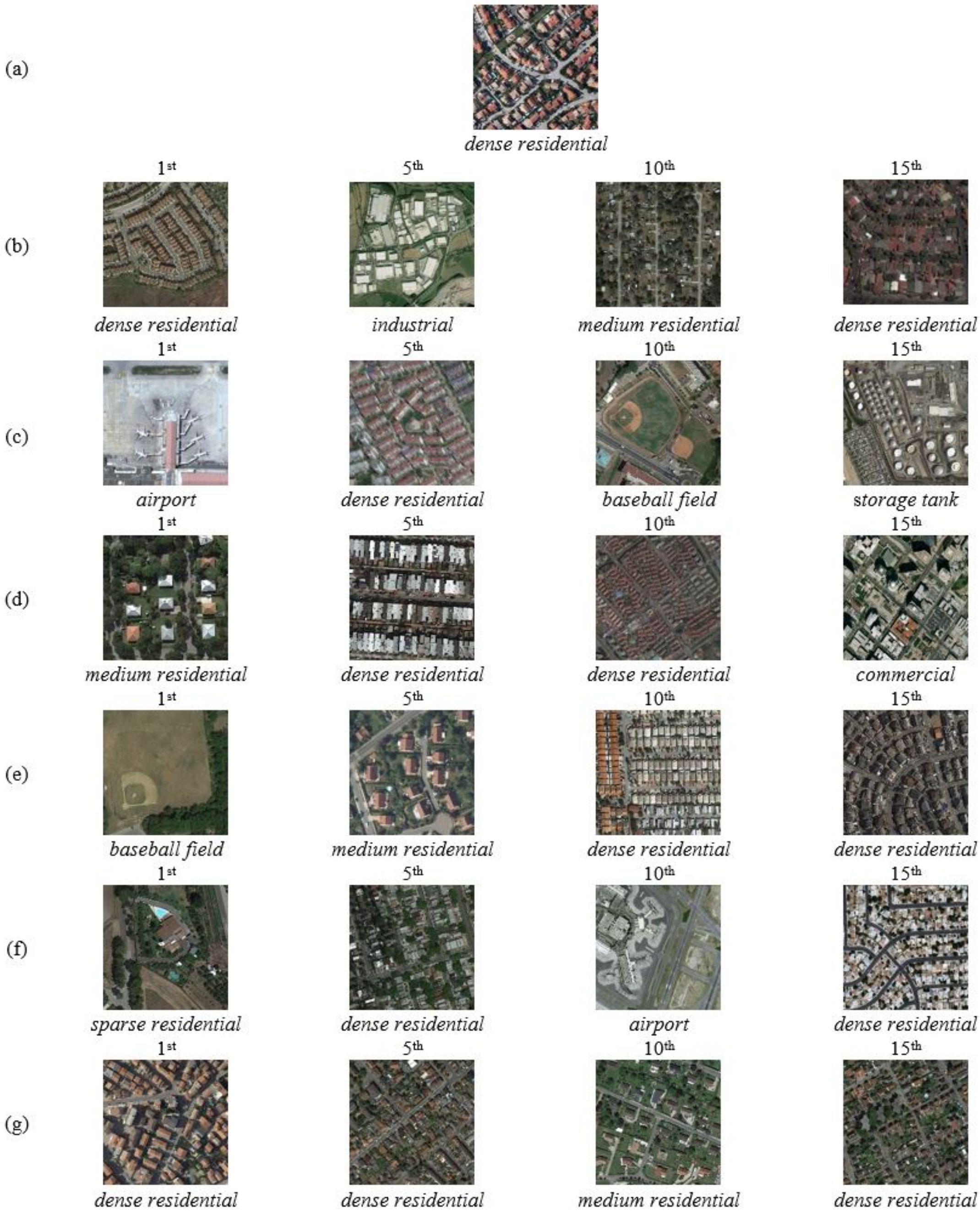
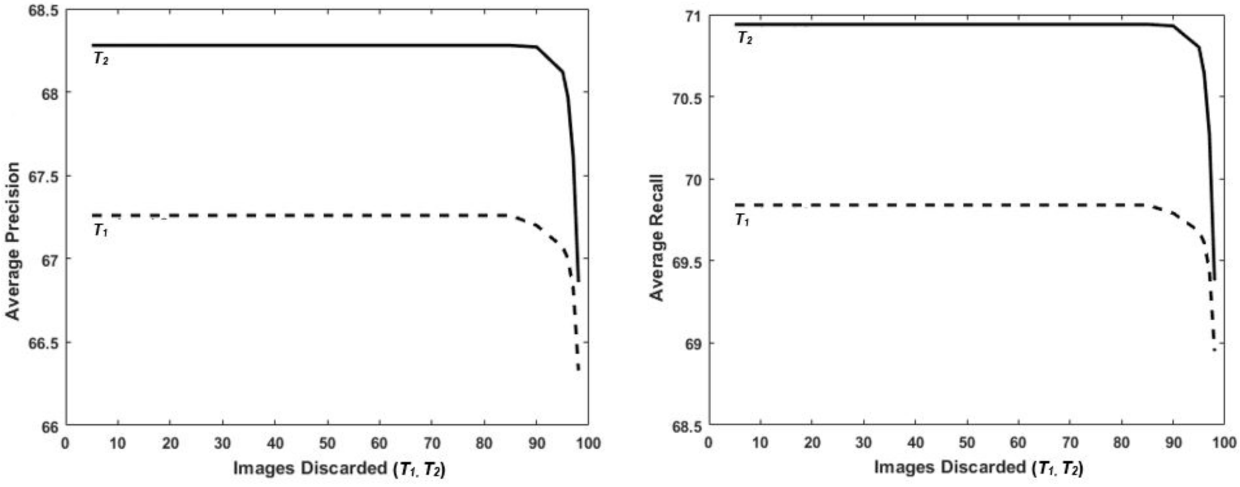


Figure 4.3: Example of (a) query image; and retrieved images by using (b) the EES descriptor, (c) the LBP descriptor, (d) the GLCM descriptor, (e) a combination of the EES and the LBP descriptors, (f) the LEH descriptor, and (g) the DMHV descriptor (AID archive).

obtained coarse to fine features in the proposed image retrieval system in the compressed domain. The implicit correspondence between the feature sets obtained between any two



(a) (b)

Figure 4.4: (a) Average precision and (b) average recall provided by the proposed PCF-CBIR system versus T_1 and T_2 (UCMERCED archive).

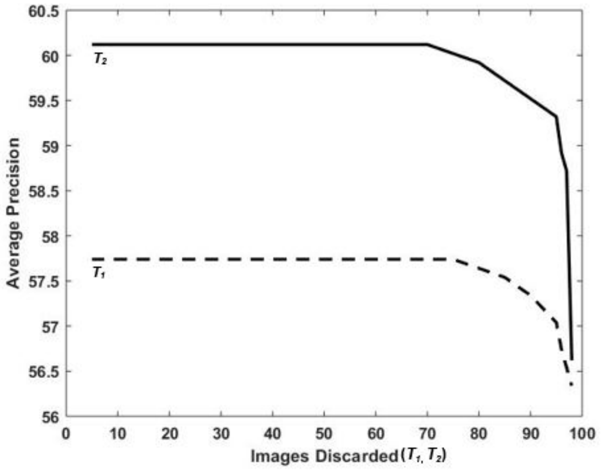


Figure 4.5: Average precision provided by the proposed PCF-CBIR system versus T_1 and T_2 (AID archive).

Table 4.4: Average precision for the proposed progressive coarse to fine RS CBIR system at each level when T_1=25% (AID archive).

Decomposition Levels	Average Precision(%)
Level 1 **(Coarsest Feature)**	55.67
Level 2 (Fine Feature)	57.74
Level 3 **(Finest Feature)**	60.12

wavelet decomposition levels adds more discriminant texture information, which is utilized to discard irrelevant images to the query image at a very early stage. In our experiments the value of the parameter T_1 that represents the percentage of images discarded at the first level is varied in the range between 0% and 100% with step-size increment of 5%.

Fig. 4.4 and 4.5 show the performance of the proposed RS CBIR system in terms of precision and recall versus the varying values of T_1 and T_2 for the UCMERCED and the AID archives, respectively. By analyzing Fig. 4.4 (UCMERCED archive), one can notice that there is no change in the performance measures when the value of T_1 varies between 0% and 90%. This shows that the compressed domain texture features obtained at a very coarse level are able to efficiently characterize the images in the archive. In other words, we can conclude that the DMHV descriptor are able to efficiently discriminate 90% of the images in the archive using only the coarser features. We can see a continuous decrease in the performance metrics when T_1>90% of the images are discarded at a very early stage using coarse features because of discarding relevant images in the initial stage. This occurs because descriptors obtained from the coarser resolution are able to characterize relevant images with respect to the query image. From the Fig. 4.5 (AID archive) we observed that there is no change in the precision values when the value of T_1 varies between 0% and 75%. This demonstrates the ability of the features obtained from the coarser level to efficiently characterize images having varying spatial resolution in the AID archive. Thus, the value of T_1 should be selected on the basis of a trade-off analysis between computational complexity and performance of the final retrieved images. On the basis of these results, we fixed the value of T_1 as 25%.

To further investigate the performance of the proposed system, we analyzed the characteristics of the images that are discarded using the coarse features and second lowest fine features. Fig. 4.6 and 4.7 show an example of the images discarded using the coarsest features and the second lowest fine features for both the archives. In detail, Fig. 4.6 demonstrates the discarded images when a query image that contains *bare soil, buildings, cars, pavement, trees* is selected from the UCMERCED archive. From the analysis, one may observe that using only the coarsest level features, one can discard highly irrelevant

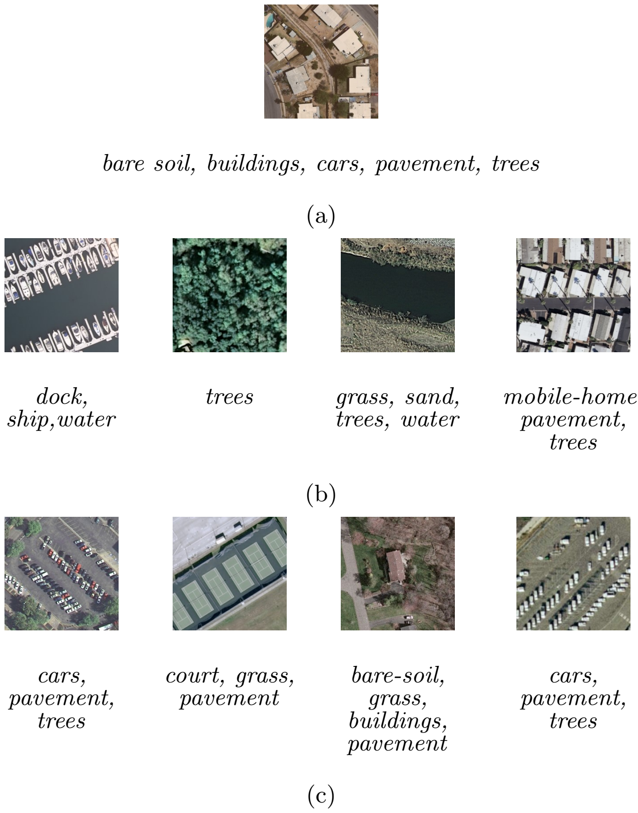
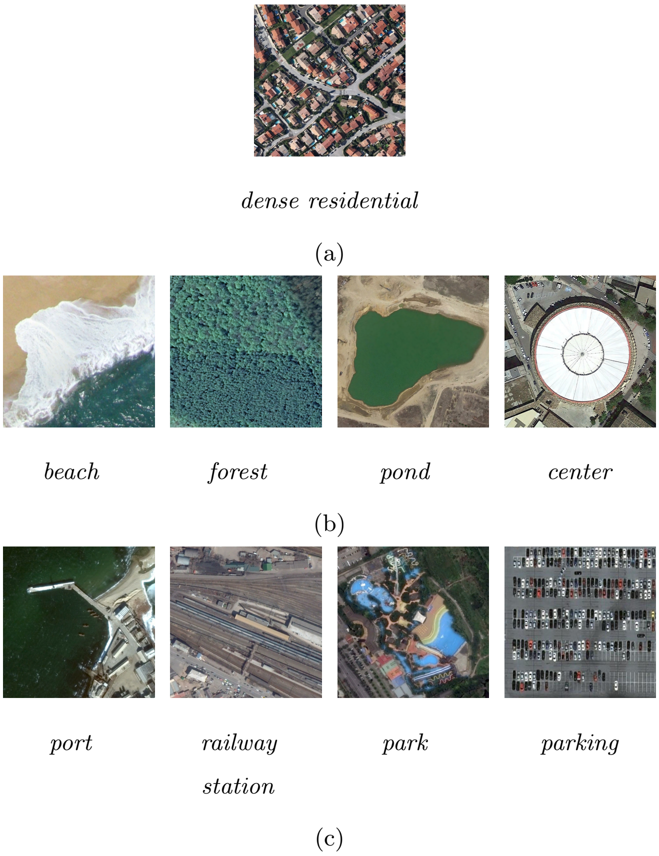

bare soil, buildings, cars, pavement, trees
(a)
dock, ship,water
trees
grass, sand, trees, water
mobile-home, pavement, trees
(b)
cars, pavement, trees
court, grass, pavement
bare-soil, grass, buildings, pavement
cars, pavement, trees
(c)

Figure 4.6: Example of (a) query image; (b) images discarded using coarsest features; and (c) images discarded using second lowest fine features (UCMERCED archive).

images to the query image that contain labels such as dock, ship, water, forest, mobile-home. This shows that the coarsest level features are enough to reject highly irrelevant images from the archive at a very early stage. This further speeds up the proposed system as only a subset of relevant images requires decoding. In the second iteration, using the fine features, the system is able to discriminate properly highly similar images with respect to the given query image. Thus, using the second lowest fine feature [see Fig. 4.6(c)] the images with more similar class label-sets such as cars, pavement, trees are discarded. We observed similar results when the AID archive is considered. Fig. 4.7 shows the images discarded when a query image is selected from *dense residential* category of the AID archive. By analyzing the figure, one can observe that using only the coarsest level features [see Fig. 4.7(b)], one can discard irrelevant images that contain labels such as *beach*, *forest*, *pond* and *center*. Using the second lowest fine features [see Fig. 4.7(c)], the proposed system is able to discriminate images that contain label sets such as *port*, *railway station*, *park* and *parking*.

Fig. 4.8 shows the behaviour of the computational time (including both decoding



dense residential

(a)

beach *forest* *pond* *center*

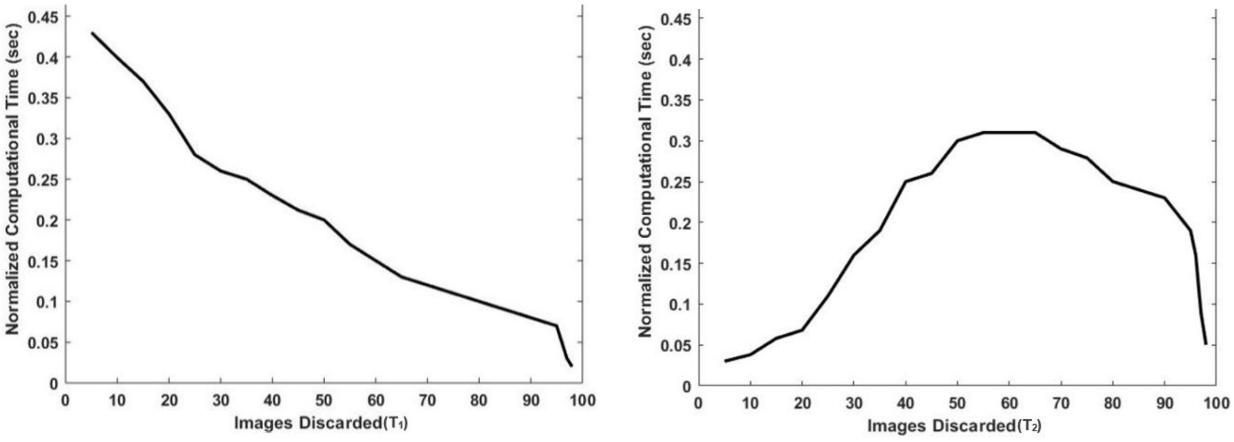
(b)

port *railway station* *park* *parking*

(c)

Figure 4.7: Example of (a) query image; (b) images discarded using coarsest features; and (c) images discarded using second lowest fine features (AID archive).

and feature extraction) versus the percentage of the images discarded after decoding the 2^{nd} wavelet decomposition level (for which T_1 percent of images are discarded) and the 1^{st} wavelet decomposition level (for which T_2 percent of images are discarded) for the UCMERCED archive. Initially, the coarse features are obtained after decoding all the N images in the archive. Then, T_1 of the images are discarded and the second lowest fine features are obtained for the subset of the remaining relevant images. Fig. 4.8-a shows the computational time required to decode and obtain features from the 2^{nd} wavelet decomposition level. From the graph, one can notice that, as the percentage of images discarded increases, the computational time required to decode and obtain features from the resulting subset of relevant images decreases. Fig. 4.8-b shows the required computational time to decode and obtain features from the 1^{st} wavelet decomposition level. When the value of T_1 decreases, the time taken to decode and obtain features after eliminating T_2 (which is defined as $1 - T_1$) also decreases and vice-versa. Thus, the graph is not linear and the computational time peaks when T_1= T_2=50%. From an analysis of



(a)

(b)

Figure 4.8: Variation in computational time (including both decoding and feature extraction) versus (a) T1 and (b) T2 (UCMERCED archive).

Table 4.5: Average precision and recall of the standard CBIR system that uses SIFT features, standard RS CBIR system without coarse to fine strategy and the proposed progressive coarse to fine RS CBIR system (UCMERCED archive).

Method	Average Precision(%)	Average Recall(%)	Decoding time (in seconds)	CBIR time (in seconds)
Standard-CBIR (SIFT features [65])	65.68	68.73	113.65	161.50
Standard-CBIR (DMHV descriptors without coarse to fine approach)	68.18	70.87	99.74	51.14
Proposed progressive-CBIR	**68.28**	**70.94**	**58.18**	**29.08**

the peaks of the graphs, we can conclude that when we reduce the number of images that require decoding to a high wavelet decomposition level, the computational time taken by the retrieval system decreases. The same behavior is also obtained when the AID archive is used.

Table 4.6: Average precision and recall of the standard CBIR system that uses SIFT features, standard RS CBIR system without coarse to fine strategy and the proposed progressive coarse to fine RS CBIR system (AID archive).

Method	Average Precision(%)	Decoding time (in seconds)	CBIR time (in seconds)
Standard-CBIR (SIFT features [65])	55.29	259.65	271.44
Standard-CBIR (DMHV descriptors without coarse to fine approach)	59.97	234.25	141.31
Proposed progressive-CBIR	**60.12**	**127.56**	**75.59**

4.4.3 Comparison of the proposed CBIR system with the state-of-the-art systems

In this subsection, we compare the effectiveness of the proposed system (proposed progressive-CBIR) with: i) a standard-CBIR system that exploits the SIFT features obtained from fully decoded images; ii) a standard-CBIR system that exploits the DMHV descriptors without coarse to fine strategy. Tables 4.5 and 4.6 report the results for the UCMERCED and the AID archives, respectively, along with the required decoding time and CBIR time. The decoding time is associated to the time required for decoding the code streams, whereas the CBIR time is associated to the time taken by both the extraction of the descriptors and the retrieval of the images. It is worth noting that in the proposed progressive-CBIR system decoding of an image from the archive depends up on its relevancy in the retrieval with respect to the given query image.

From the tables, one can observe that the proposed progressive-CBIR system provides higher accuracies with significantly reduced decoding and CBIR times for both archives compared to the standard-CBIR system that uses SIFT features. As an example, the proposed system outperforms the standard CBIR system by almost 3% in precision and 2% in recall for the UCMERCED archive, and almost 4% in average precision for the AID archive. The accuracies obtained by using the standard-CBIR system that exploits the same descriptor without applying the proposed coarse to fine strategy are very similar to those obtained by the proposed system for both archives. However, required decoding and CBIR times for the proposed progressive-CBIR system are almost half of the time

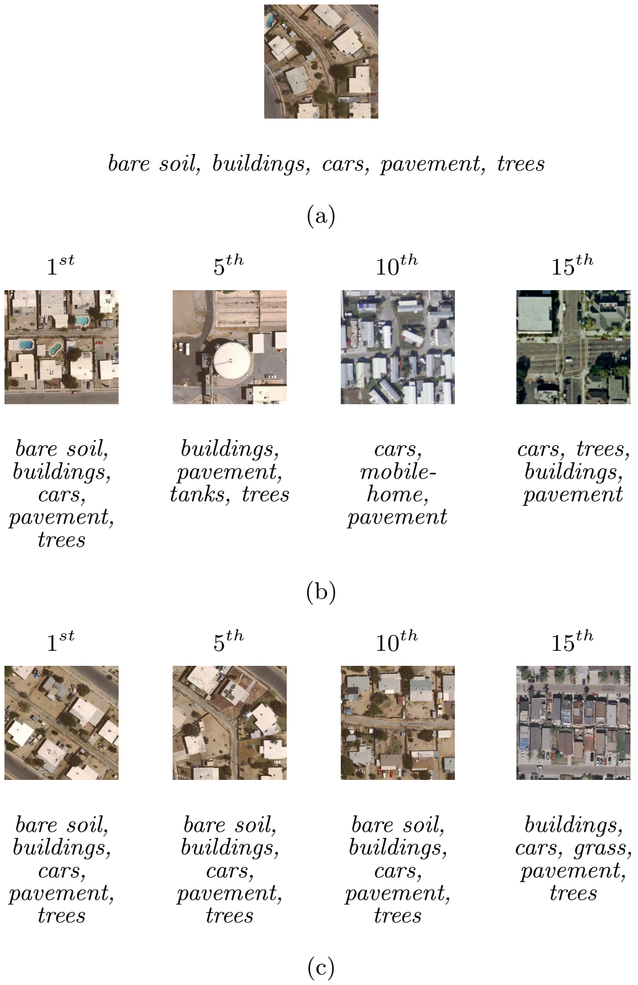
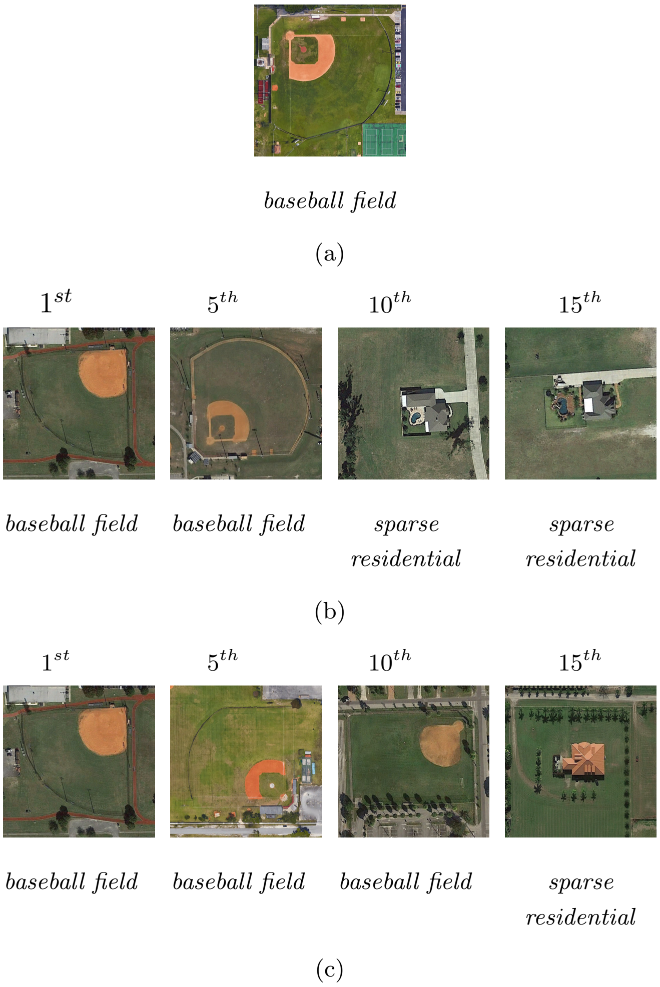

bare soil, buildings, cars, pavement, trees
(a)
1st
5th
10th
15th
bare soil, buildings, cars, pavement, trees
buildings, pavement, tanks, trees
cars, mobile-home, pavement
cars, trees, buildings, pavement
(b)
1st
5th
10th
15th
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, pavement, trees
bare soil, buildings, cars, pavement, trees
buildings, cars, grass, pavement, trees
(c)

Figure 4.9: Example of (a) query image; and retrieved images by using (b) the standard-CBIR system that uses SIFT features; and (c) the proposed progressive-CBIR system (UCMERCED archive).

required for the standard systems. In detail, for the UCMERCED archive the standard RS CBIR system that uses SIFT features (which requires complete decoding of the images) takes 113.65 seconds as the decoding time, whereas the proposed system takes 58.18 seconds. This shows that there is a sharp improvement in computational time (with same performance measures as the standard-CBIR system) when the image retrieval is performed in the compressed domain. All these results confirm that in the proposed system, discarding irrelevant images and adopting a progressive coarse to fine strategy shows significant improvements over the existing RS CBIR systems. Note that, as shown in the tables, discarding irrelevant images at a very early stage considerably reduces the CBIR time. Fig. 4.9 shows an example of results with a query image selected from the UCMERCED archive that includes six class-labels: *bare soil*, *buildings*, *cars*, *pavement*



(a)

(b)

(c)

Figure 4.10: Example of (a) query image; and retrieved images by using (b) the standard system without coarse to fine strategy; and (c) retrieved images by proposed progressive-CBIR system (AID archive).

and *tree*, while Fig. 4.10 shows an example of results with a query image that belongs to the *baseball field* category within the AID archive. Through these examples one can see that the images retrieved from the progressive-CBIR are more relevant than those retrieved by the standard-CBIR system that uses SIFT features for both archives. As an example, the images retrieved using standard-CBIR system based on SIFT features does not include most of the class label sets as that of the query image for the UCMERCED archive (Fig. 4.9).

4.5 Conclusion

In this chapter we have introduced a novel content-based image retrieval (CBIR) system that accomplishes a coarse to fine progressive RS image description and retrieval in partially decoded JPEG 2000 compressed domain. The proposed system considers that the amount of data that needs to be entropy decoded is directly related to the relevancy of the images in the retrieval process. To reduce the time required for fully-decoding images, the proposed system initially decodes only the code-blocks associated to the lowest wavelet resolution of all images in the archive. Then, based on the similarities estimated by the histogram intersection kernel among the coarse resolution wavelet descriptors of the query image and those of the archive images, the most irrelevant images related to the smallest similarity values are discarded. This step allows identification and elimination of the most irrelevant images at a very early stage to reduce the subsequent decoding time. The processes of code-blocks decoding and elimination of the irrelevant images (with respect to the similarities among the descriptors associated to the considered wavelet resolutions) are iterated until the code streams associated to the highest wavelet resolution are decoded. Then, the most similar images to the query are selected. By this way, the proposed system exploits a multiresolution and hierarchical feature space representation and accomplishes a progressive RS CBIR with significantly reduced retrieval time. To characterize each resolution level, a texture descriptor that models the distribution of moduli of the horizontal and vertical detail coefficients is used. In order to evaluate the similarities among the descriptors that model different wavelet resolutions, the pyramid match kernel is exploited. The pyramid match kernel computes the weighted sum of all the implicit correspondences between the descriptors of the different wavelet decomposition levels by considering the importance of the descriptors at different wavelet resolution levels.

Experimental results obtained on a benchmark archive show that the proposed system results in similar accuracies with respect to a standard-CBIR system (which operates on the fully decoded image domain) with significantly reduced decoding and thus retrieval time. This is due to the progressive removal of a very large amount of irrelevant images, which allows to apply the final retrieval process only to a very small set of images (which are highly relevant to the query image). We emphasize that this is a very important advantage, because the main objective of large-scale CBIR is to optimize the search and retrieval time with a minimum amount of fully decoded images. Thus, the proposed system is promising for possible operational applications due to both its general properties and also its simplicity in the implementation. Note that the archives used in the experiments are benchmarks. However, in many real applications the search is expected to be applied to much larger archives. For large scale CBIR problems, by using our system the gain in both retrieval and decoding time is expected to be increased considerably with respect to the standard-CBIR systems. As a final remark, we point out that the proposed system can be easily adapted to the CBIR problems for which images are compressed by other compression algorithms by properly defining the image description algorithm in the (partially) compressed domain.

Chapter 5

Approximating Wavelet Representations through Deep Neural Network for Compressed Remote Sensing Image Scene Classification

In this chapter we propose a novel approach to achieve scene classification in the JPEG 2000 compressed RS images. The proposed approach consists of two main steps: i) approximate finer resolution sub-bands of reversible biorthogonal wavelet filters used in JPEG 2000; and ii) characterize high-level semantic content of the approximated wavelet sub-bands and perform scene classification based on the learnt descriptors. This is achieved by taking as input codestreams associated with the coarsest resolution wavelet sub-band to approximate finer resolution sub-bands using a number of transposed convolutional layers. Then, a series of convolutional layers is used to model the high-level semantic content of the approximated wavelet sub-band. Thus, the proposed approach models the multiresolution representation of the JPEG 2000 compression algorithm in an end-to-end trainable unified neural network. In the classification stage, the proposed approach takes as input only the coarsest resolution wavelet sub-bands, thereby reducing the time required to apply decoding. Experimental results obtained on two benchmark aerial image archives demonstrate that when compared to traditional RS scene classification approaches (which requires full image decompression), the proposed method significantly reduces the computational time keeping similar classification accuracies.

5.1 Introduction

Due to the recent advances in satellite technology, RS society has witnessed a huge explosion in the volume of the data. Simultaneously, the amount of insightful information that can be extracted from them has also increased. One of the challenging issues faced by the RS society is the development of efficient scene classification approaches in real large-scale RS image archives. Although pixel-level and object-based classification has demonstrated excellent performance, they do no consider the high-level semantic information within the images. To address this, remarkable efforts are carried out to develop efficient scene classification approaches. Scene classification methods assign class labels to each image and several efforts have been made to develop effective approaches over the past several years. Performance of any scene classification method mainly depends on obtaining powerful discriminative features from the images. Conventional methods that obtain handcrafted traditional descriptors to perform scene classification are extensively time demanding and computationally complex. In the recent years, the potential of DL methods has gained increasing attention to address RS scene classification problems due to its remarkable ability to learn discriminative image representations [83; 84; 85; 86; 98; 99; 100; 101]. Their ability to learn high-level semantic content of the images has resulted in obtaining high classification performance over the state-of-the-art traditional methods (see Section 3.2). Although all the existing DL models have shown excellent performance in RS scene classification problem, they require fully decompressed input images.

To address this limitation, in this chapter we present a novel approach that benefits from DNN to perform scene classification by minimizing the amount of image decompression. The proposed approach includes two main steps: (i) approximating wavelet sub-bands or fully decoded image; and (ii) feature extraction and classification of the approximated wavelet sub-band. The proposed approach initially approximates finer (highest) wavelet resolution sub-bands of the reversible biorthogonal filter used in the JPEG 2000 from the coarsest (lowest) resolution wavelet sub-band. To achieve this, the proposed approach employ a series of deconvolutional layers through which the finer resolution wavelet sub-bands (approximated image) are approximated. Then, the high-level semantic content of the approximated wavelet sub-bands (approximated mage) are learned through a sequence of convolutional layers and finally performs scene classification. By this way, the proposed approach utilizes the multiresolution paradigm inherent within the JPEG 2000 compression algorithm to achieve an efficient scene classification at a faster computational rate. Experimental results performed on two benchmark archives demonstrate the effectiveness of the proposed method.

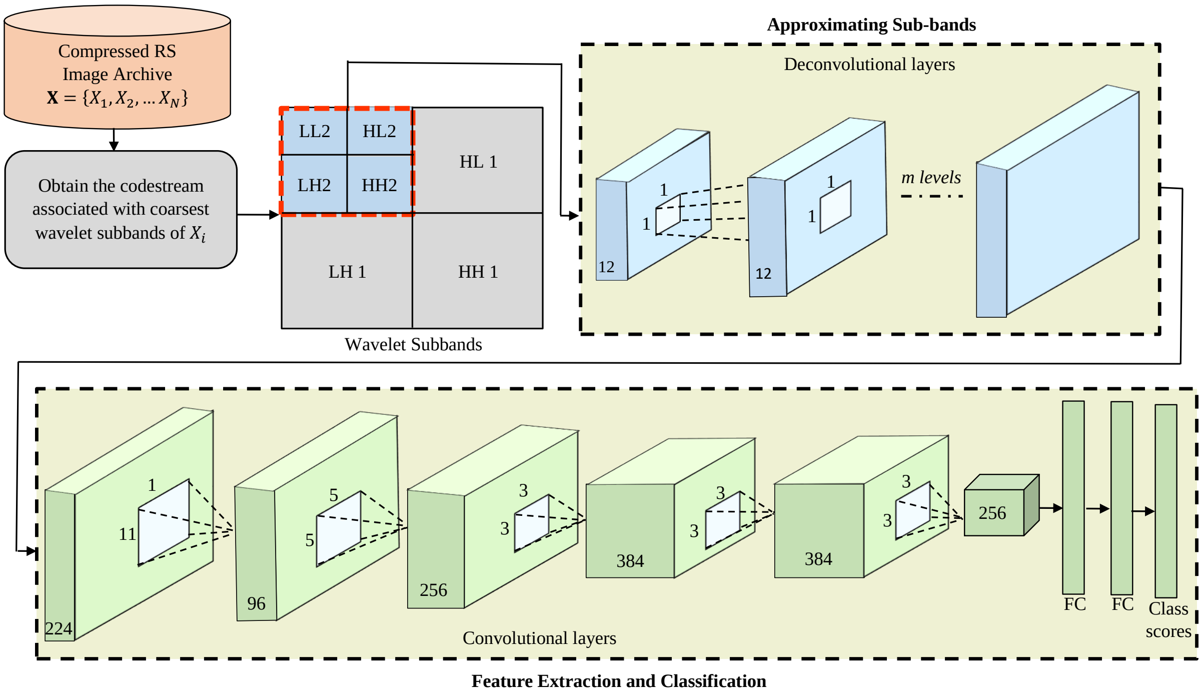


Figure 5.1: Block scheme of the proposed approximation and classification approach in the compressed domain.

5.2 Proposed Scene Classification Approach in JPEG 2000 Compressed Domain

5.2.1 Problem Formulation

Let $\mathbf{X} = \{X_i\}_{i=1}^{N}$ be an archive that contains N JPEG 2000 compressed RS images, where X_i represents the i^{th} image. The main objective of the proposed method is to assign to a given input image $X_i \in \mathbf{X}$ a class label $y_i \in \mathbf{Y}$, where $\mathbf{Y}$ is a set of $\mathbf{Q}$ class labels. Let us assume that all the images in the archive are decomposed up to L resolutions. Each image in the archive will be represented as one approximation sub-band and $3L$ detail sub-bands (i.e. horizontal, vertical and diagonal). In JPEG 2000 compressed image archive, the straightforward approach to perform scene classification is to: i) apply entropy decoding to the codestreams associated with all the images in the archive; and ii) obtain the image descriptors. However, decoding all the images from a compressed archive is time demanding and computationally expensive. Thus, we propose a novel approach to achieve scene classification in the JPEG 2000 compressed domain that benefits from DNNs. Figure 5.1 shows the block scheme of the proposed approach.

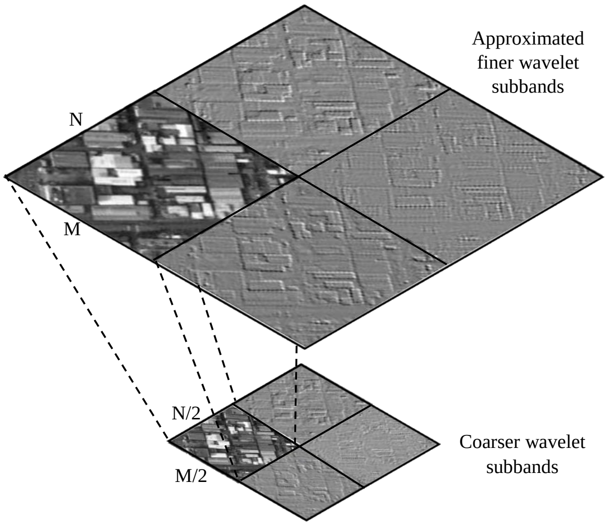


Figure 5.2: Example of approximation of a finer level wavelet sub-bands using transposed convolution.

5.2.2 Approximation of the wavelet coefficients

We propose a novel approach based on DNN that efficiently approximates a decompressed image to perform scene classification in a large scale JPEG 2000 compressed image archive. Our objective is: i) to improve the computational time when compared to models that require fully decoded images; and ii) to implement a novel DL model that performs scene classification in the compressed domain with minimal decompression. To achieve this, the proposed approach initially obtains the codestream associated with the coarsest level wavelet sub-band (see Fig. 5.1) that provides the global scale information of any given image. Accordingly, the proposed approach approximates the finer level (higher resolution) wavelet sub-bands (or the image itself) through a series of transposed convolutional layers. The approximated finer level wavelet sub-band (image) provides detailed information of a given image. To achieve this, the proposed approach considers m transposed convolutional layers, where m corresponds to the number of wavelet decomposition levels that were initially used to compress a given image $X_i \in \mathbf{X}$. We obtain the features associated with the approximated wavelet sub-band (partially decompressed or image level information) by using five convolutional layers and two fully connected (FC) layers to obtain the classification scores. During classification, the proposed approach requires only the coarsest level wavelet sub-band information thereby reducing the amount of time required to perform decompression of images in the archive. The detailed information regarding the approximating wavelet sub-band and feature extraction are provided below.

Given a JPEG 2000 compressed image X_i, we initially decode the k-th codestream (where $k << m$) associated with the given image to obtain the approximated and detail wavelet sub-bands where m represents the total number of wavelet decomposition levels used to compress the images in the considered archive. Let $G^L = \{a^L_{X_i}, h^L_{X_i}, v^L_{X_i}, d^L_{X_i}\}$ denote the approximation, horizontal, vertical and diagonal sub-bands of an image X_i at the L^{th} wavelet decomposition level (coarsest wavelet sub-band). Let $A^{L-1} = \{a^{L-1}_{X_i}, h^{L-1}_{X_i}, v^{L-1}_{X_i}, d^{L-1}_{X_i}\}$ be the next finer level approximated sub-bands at level $L-1$. The proposed approach initially approximates the finer level wavelet sub-bands that are estimated from the coarsest (low-level) wavelet sub-bands to learn the high-level semantic contents of the approximated finer level wavelet sub-band or image. To achieve this, k-th codestream associated with the compressed image X_i are modelled to approximate the m-th level wavelet sub-bands. In the proposed model, approximation is performed using a series of transposed convolutional layers. In CNN, the convolution and pooling operation generally reduces the size of the output. Thus, we can consider, *'convolution'* as a matrix multiplication between the given input A^{L-1} and $\mathbf{C}$ to obtain G^L, where $\mathbf{C}$ represents the sparse matrix which can be obtained as:

$$G^L = \mathbf{C} \cdot A^{L-1} \tag{5.1}$$

The non-zero elements in the sparse matrix C can be constructed using the kernel coefficients of the convolution operation as follows:

$$C = \begin{bmatrix} k_{11} & ... & k_{1q} & 0 & ... & k_{2q} & ... & k_{pq} & 0 & ... \\ 0 & k_{11} & ... & k_{1q} & 0 & ... & k_{2q} & ... & k_{pq} & ... \\ 0 & 0 & k_{11} & ... & k_{1q} & 0 & ... & k_{2q} & & ... \\ ... & & & & & & & & & ... \\ 0 & 0 & 0 & 0 & & & ... & & & k_{pq} \end{bmatrix}, \tag{5.2}$$

where p and q represent the kernel size and k_{ij} is the element of the kernel (where i and j are the row and column indices of the kernel, respectively). Convolution operation takes the input matrix A^{L-1} which is then flattened into a vector and multiplies the flattened input with C. The matrix multiplication result is reshaped to obtain the final output G^L. It is worth noting that during the forward and backward passes of CNNs, convolution operations are applied with C and C^T, respectively.

In computer vision and pattern recognition, *transposed convolution* has proved to be an efficient algorithm that uses the gradient of the convolution operation (for a given image) to perform image restoration and reconstruction [119]. Our proposed approach approximates finer wavelet sub-bands using transposed convolution as shown in Fig. 5.2. Given a kernel k, the transposed convolution multiplies the flattened input vector G^L with $\boldsymbol{C^T}$ during the forward pass and multiplies $\boldsymbol{(C^T)^T = C}$ during the backward pass to obtain A^{L-1}. The finer level wavelet sub-bands can be obtained as:

$$A^{L-1} = \mathbf{C}^T \cdot G^L \tag{5.3}$$

Thus, in this operation we swap the backward and forward pass of the convolution operation which is used in a standard CNN networks. Accordingly, using m transposed convolutional layers, the proposed approach allows to approximate the image A^m. For the transposed convolutional layers, if we use a stride S, padding P and kernel k, then the size of the approximated wavelet sub-bands (A^{L-1}_{size}) obtained from the coarser level wavelet sub-band (A^{L}_{size}) is given by:

$$A^{L-1}_{size} = S * (A^{L}_{size} - 1) + k - 2P. \tag{5.4}$$

The proposed approximation approach reflects the inherent multiresolution paradigm within the JPEG 2000 compression algorithm within an end-to-end unified framework. While approximating sub-bands, we consider two scenarios:

Scenario 1: Minimal Decoding

In this Scenario, the proposed approach uses only the codestreams associated with the coarsest level (L^{th} level) wavelet sub-bands to approximate the finer level sub-bands (image itself). Here, the aim is to minimize the amount of decompression time required to perform scene classification by approximating wavelet sub-band (image) using only the coarsest level sub-bands. The coarsest level wavelet sub-band provides global scale information of the considered image. Thus, here, although the amount of time required for decompression is significantly reduced, the quality of approximation is moderately diminished. Fig. 5.3 illustrates the case when the proposed approach takes the codestreams associated with 32 × 32 coarsest level wavelet sub-band to approximate the image A^m using m transposed convolutional layers.

Scenario 2: Partial Decoding

In this Scenario, the proposed approach takes the coarsest level (L^{th} level) wavelet sub-band information to decode the finer level ($L - 1^{th}$ level) wavelet sub-band, which is exploited to approximate the finest level wavelet sub-bands (image itself). Here, the amount of decompression time required is reduced moderately to achieve favourable performance, when compared to the case where the images requires full decompression. The finer level wavelet sub-bands provide fine scale information of a given image. Thus, the wavelet sub-bands (image) approximated from the finer level sub-bands incorporate the detailed fine scale information that enhance the classification accuracy with moderate

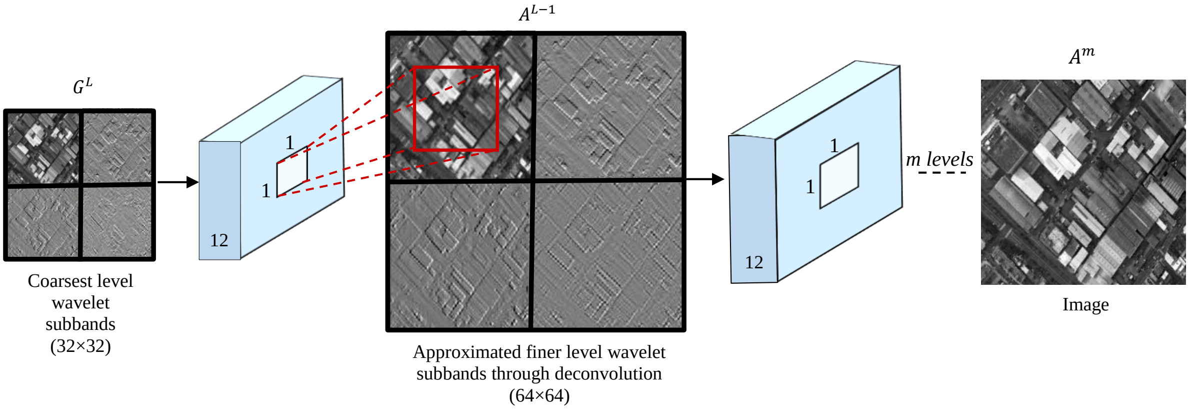

G^L
Coarsest level wavelet subbands (32×32)
1
1
12
A^{L-1}
Approximated finer level wavelet subbands through deconvolution (64×64)
1
1
12
m levels
A^m
Image

Figure 5.3: Illustration of the proposed approach when the approximated image is obtained from codestream of coarsest level wavelet sub-band.

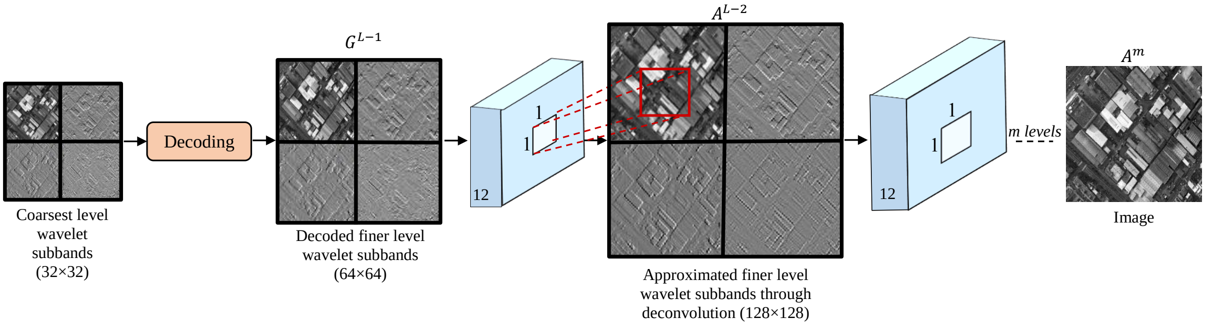

Coarsest level wavelet subbands (32×32)
Decoding
G^{L-1}
Decoded finer level wavelet subbands (64×64)
1
1
12
A^{L-2}
Approximated finer level wavelet subbands through deconvolution (128×128)
1
1
12
m levels
A^m
Image

Figure 5.4: Illustration of the proposed approach when the approximated image is obtained from codestream of decoded finer level wavelet sub-band.

reduction in time. Fig. 5.4 illustrates the case when a 32 × 32 coarsest level wavelet sub-band is employed to decode the finer level sub-band of size 64 × 64. Then, deconvolution is applied to the decoded finer level wavelet sub-band to approximate the finest level wavelet sub-band (image).

5.2.3 Feature Extraction and Classification

The feature extraction and classification step aims to obtain features from the approximated wavelet sub-bands (image). To this end, we consider a model with five convolutional layers with a number of filters similar to that of the AlexNet [120] and two fully connected (FC) layers. By modifying the feature extraction and classification steps, we can obtain powerful discriminative features. To demonstrate the effectiveness of recent DL models when used in the compressed domain wavelet subband information, we selected the ResNet50 [121] architecture to compare with the results obtained by the AlexNet.

Then, the output obtained from the final FC layer is mapped into **Q** classification scores. To reduce information loss, we considered zero padding and stride of 1 in each convolutional layer, which is followed by a max-pooling layer except the third and fourth. As no pretrained model on wavelet coefficients are available, the proposed end-to-end model was trained from scratch with random weight initialization. The final classification of the model is obtained by learning the approximations obtained through several transposed convolutional layers. The total loss function ($\mathcal{L}_{total}$) of the proposed approach is the sum of the approximation loss ($\mathcal{L}_{approximation}$) and classification loss ($\mathcal{L}_{classification}$), which is obtained as:

$$\mathcal{L}_{total} = \mathcal{L}_{classification} + \mathcal{L}_{approximation} \tag{5.5}$$

The $\mathcal{L}_{approximation}$ function is obtained as the sum of mean squared error (MSE) between the approximated wavelet sub-bands and decoded wavelet sub-bands at each level l which is obtained as:

$$\mathcal{L}_{approximation} = \sum_{i=L}^{1} \sum_{j=1}^{m} \sum_{k=1}^{n} ||A^i(t[j,k]) - D^i(t[j,k])||^2 \tag{5.6}$$

where $m \times n$ represents the size of the considered wavelet sub-bands at level l, $t[j,k]$ denote the wavelet coefficient at position $[j,k]$ and D^i represents the decoded wavelet sub-band at any given level i.

To evaluate the $\mathcal{L}_{classification}$, we chose the cross-entropy loss function, which is predominantly used for scene classification problems and is defined as:

$$\mathcal{L}_{classification} = -\sum_{i=1}^{Q} y_i log\hat{y}_i \tag{5.7}$$

where $\hat{y}_i$ denotes the predicted class label. To improve the performance batch normalization (BN), dropout was carried out after each convolutional layer. To overcome vanishing gradient problem, Rectified Linear Unit (ReLU) activation was used after both the convolutional and transposed convolutional layers. Section 5.3 provides the more detailed information regarding the training details and the parameters. It is worth noting that, the proposed end-to-end model can be adopted to perform scene classification, where the images are compressed using JPEG 2000 compression algorithm as well as when the images are compressed using any wavelet based approach.

5.3 Dataset Description and Experimental Setup

Several experiments were performed to evaluate the performance of the proposed approach on two benchmark archives. The first one is the NWPU-RESISC45 benchmark archive

Table 5.1: Number of images considerd for each archive in the training, validation and test data.

Image Archive	Training	Validation	Test
NWPU-RESISC45	25200	3150	3150
AID	8000	1000	1000

that consists of 31,500 single-labeled images with 45 different categories (i.e. airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station and wetland). Each category has 700 scene classes and each image in the archive has size 256×256 with a varying spatial resolution between 0.2m to 30m per pixel. The reader is referred to [43] for detailed information.

The second archive is the AID benchmark archive that contain 10,000 single-labeled high-resolution RS images with 30 different categories (i.e. airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, viaduct). Each image has size 600×600 pixels with a spatial resolution in the range from 0.5m to 8m. For more detailed information about the image archive reader is referred to [37].

To assess the effectiveness of the proposed model, the images of both the archives were compressed using the JPEG 2000 algorithm. Due to the minimum codeblock size constraint (see Section 2.1), we considered a three level wavelet decomposition based on the size of the images for both the archives ($L = 3$). The codestreams associated with the coarsest wavelet sub-band ($l = 3$) is used as the input to the proposed approach. The number of transposed convolutional layers (m) is equivalent to the number of wavelet decomposition levels used in the considered image archive. To avoid information loss, we selected the size of the filter as 1×1 with stride 1 and padding 0. The number of filters used for approximating the wavelet sub-band is 12×12 and the image is 3×3. During the case study of Scenario 2, we considered decoding upto $(m-1)$ wavelet decomposition levels. Both the labeled image archives were initially divided into three subsets: training (80%), validation (10%) and test (10%) as shown in Table I. Images included in each subset were randomly sampled. The training of the proposed model was carried out with the Stochastic Gradient Descent (SGD) that uses the Adaptive Moment Estimation (Adam). During training, the Xavier initialization method was used to learn the weights and parameters for the proposed model. As there are no pretrained models to perform scene classification in the compressed domain, that use wavelet coefficients, all the experiments were performed starting from scratch. In addition, to achieve accurate performance, experiments were carried out varying learning rate between 0.1 and 0.0001. The performance of the proposed architecture was assessed quantitatively and qualitatively by using classification accuracy, training time (in sec), validation time (in sec), test time (in sec) and Root Mean Square Error (RMSE) of the approximated sub-band images. The performance of the proposed architecture was assessed quantitatively and qualitatively by using: 1) classification accuracy; 2) computational time (in sec) of training, validation and test phases; and 3) Root Mean Square Error (RMSE) of the approximated sub-band images. It is worth noting that computational time of the test phase was considered as classification time. All the experiments were performed in Nvidia Tesla V100.

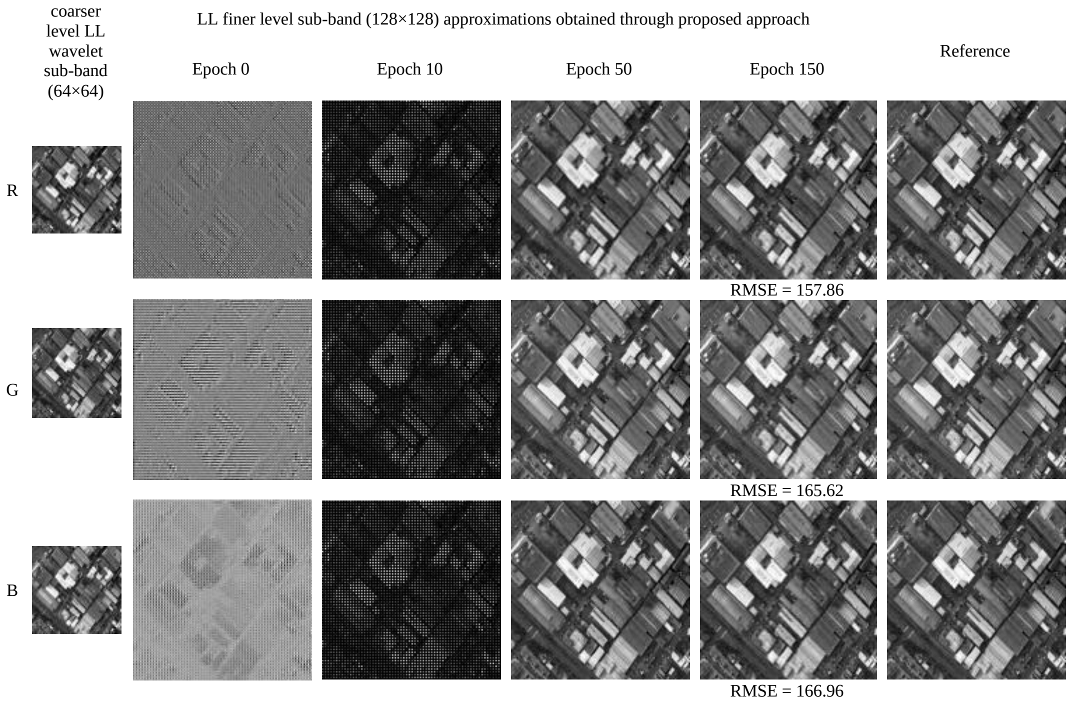

coarser level LL wavelet sub-band (64×64)
LL finer level sub-band (128×128) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 157.86
RMSE = 165.62
RMSE = 166.96

Figure 5.5: Qualitative results of sub-band approximations associated to LL wavelet sub-band of an image belonging to building category in the NWPU-RESISC45 archive.

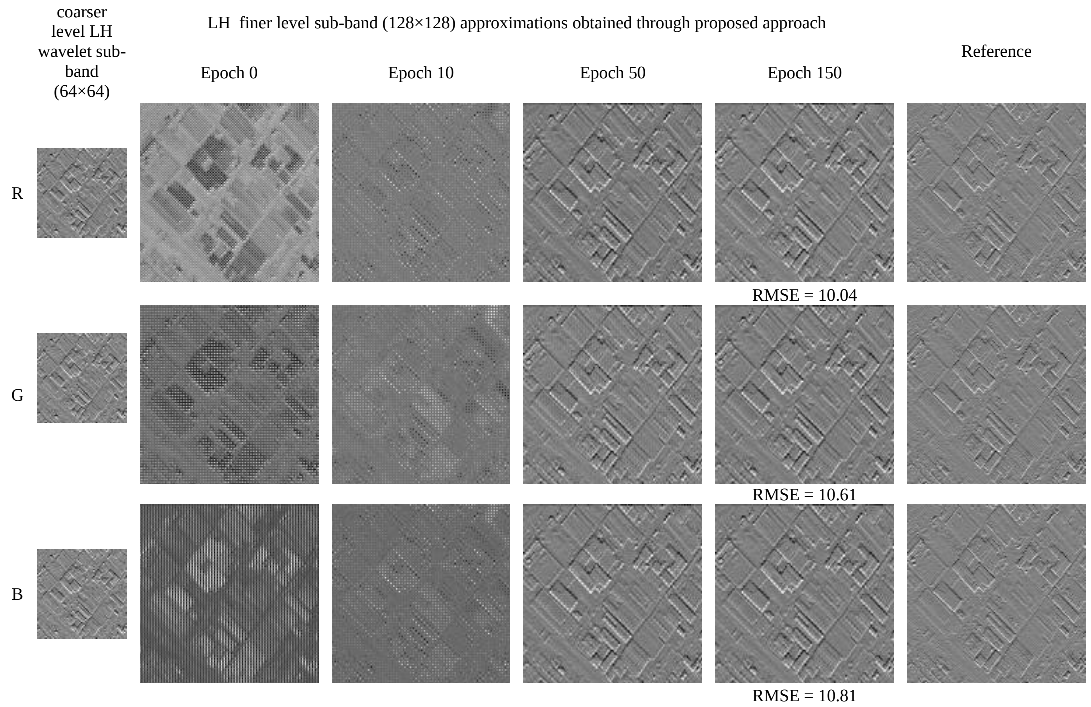

coarser level LH wavelet sub-band (64×64)
LH finer level sub-band (128×128) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 10.04
RMSE = 10.61
RMSE = 10.81

Figure 5.6: Qualitative results of sub-band approximations associated to LH wavelet sub-band of an image belonging to building category in the NWPU-RESISC45 archive.

5.4 Experimental Results

To evaluate the effectiveness of the proposed approach, we performed several experiments to: i) assess the quality of the proposed approximated images compared to the decoded wavelet sub-band (image) ones; ii) analyze the performance of the proposed approximation approach for Scenarios 1 and 2 (mentioned in Section 5.2.2); and iii) compare the performance and computational gain with respect to a standard CNN. In the first set of experiments, we assess the qualitative as well as quantitative performances of the proposed approximation approach for scene classification for both NWPU-RESISC45 and AID benchmark archives.

5.4.1 Qualitative Analysis of the Approximated Images

This subsection provides the analysis of the images obtained from the proposed approximation approach for both NWPU-RESISC45 and AID archive. To this end, we considered two different cases under each Scenario 1 and 2 where:

1. Scenario 1 - the coarsest level wavelet sub-bands are used to approximate the image level information;

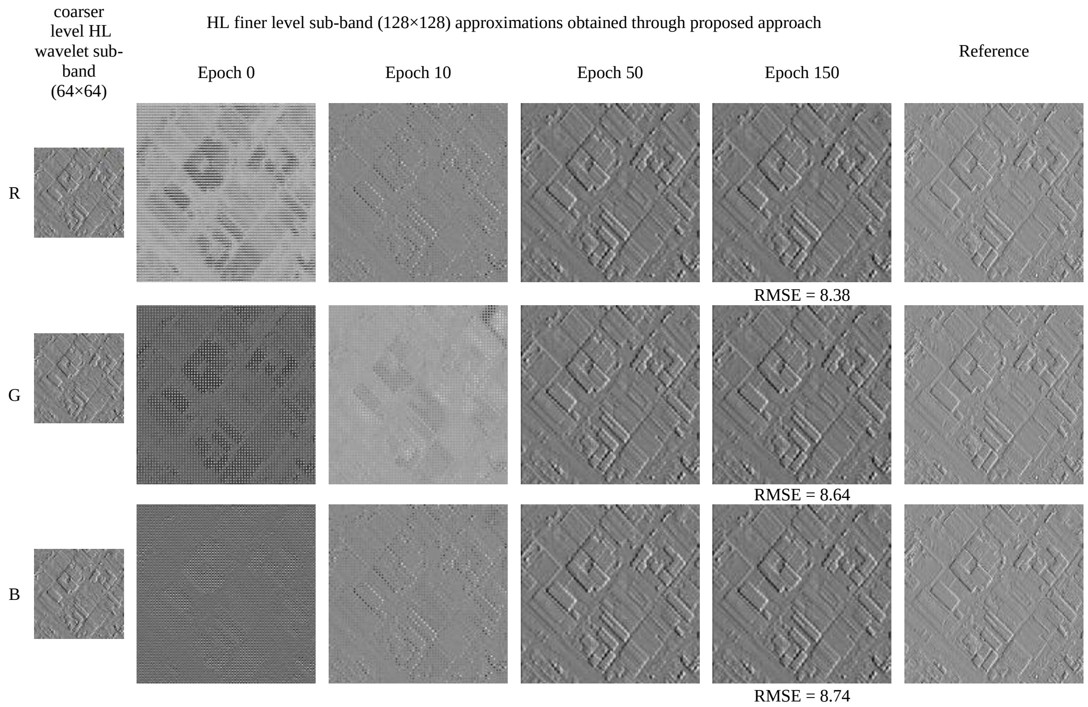

coarser level HL wavelet sub-band (64×64)
HL finer level sub-band (128×128) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 8.38
RMSE = 8.64
RMSE = 8.74

Figure 5.7: Qualitative results of sub-band approximations associated to HL wavelet sub-band of an image belonging to building category in the NWPU-RESISC45 archive.

2. Scenario 1 - the coarsest level wavelet sub-bands are used to approximate the intermediate finer level wavelet sub-bands;

3. Scenario 2 - decoded finer level wavelet sub-bands are used to approximate the image level information;

4. Scenario 2 - decoded finer level wavelet sub-bands are used to approximate intermediate finest level wavelet sub-bands.

Fig. 5.5-5.12 show the approximated images obtained for LL, LH, HL and HH wavelet sub-bands for the NWPU-RESISC45 and AID archive building category when the experiments were performed with AlexNet architecture. To qualitatively analyze the efficiency of the proposed approach, we provide the RMSE value between the approximated image and the decoded image. Given a coarser level wavelet sub-band (64 × 64) from the NWPU-RESISC45 archive building category, one can notice that the proposed approach is efficient to model the finer level wavelet sub-band (128 × 128). It converge fast (around Epoch 50) for all the wavelet sub-bands. Furthermore, we can also observe that although the approximated image is slightly blurred, the model is able to learn the semantic content of the approximated image of the building category. The RMSE values obtained for LL sub-

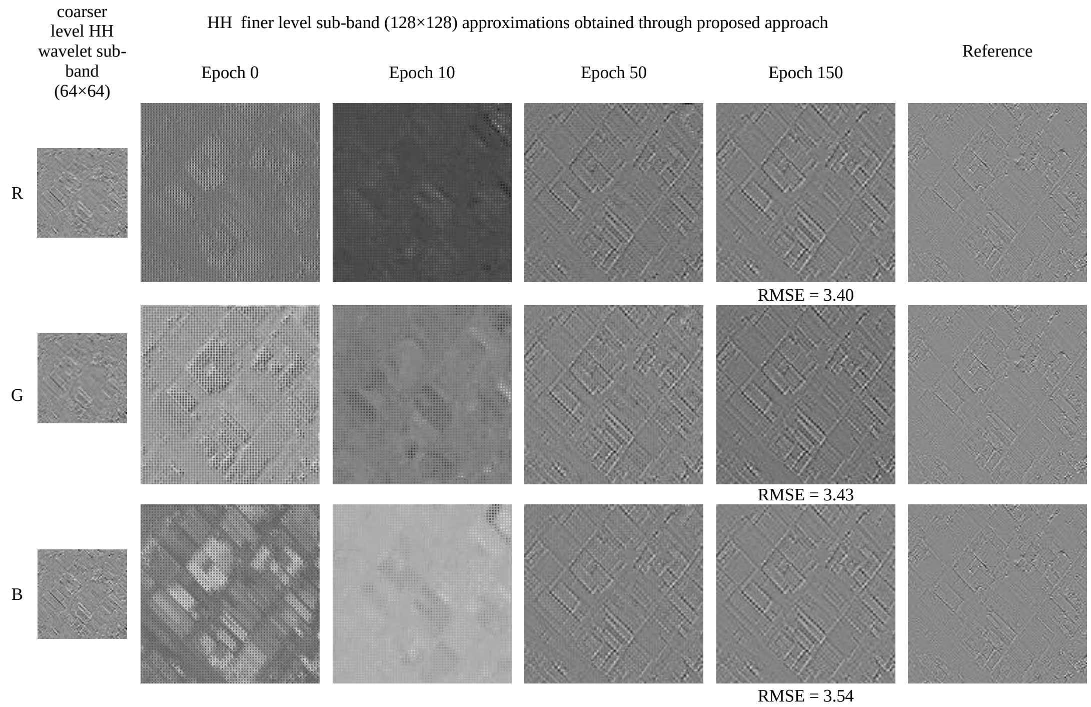

coarser level HH wavelet sub-band (64×64)
HH finer level sub-band (128×128) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 3.40
RMSE = 3.43
RMSE = 3.54

Figure 5.8: Qualitative results of sub-band approximations associated to HH wavelet sub-band of an image belonging to building category in the NWPU-RESISC45 archive.

bands are 157.86, 165.62, 166.96 for Red (R), Green (G) and Blue (B) bands, respectively. Transposed convolution used to approximate the finer level wavelet sub-bands (image) introduces a loss to the fine-scale detailed information. This is visible from the LL sub-band finest approximated images of (128 × 128) (see Fig. 5.5). The RMSE values for the HL (vertical) sub-bands which are 8.38, 8.64 and 8.74 for RGB bands, respectively. In addition, we also notice decreased RMSE values for the detail wavelet sub-bands (which are LH, HL and HH) when compared to the approximation wavelet sub-band (which is LL). Thus, we can see that the transposed convolution used efficiently approximates the detail wavelet sub-bands. Given a coarser wavelet sub-band (150 × 150) from the AID archive railway station category, we can observe that the proposed approach starts converging around Epoch 50 and that it models the edge information modelled effectively. The RMSE values obtained for the LL wavelet sub-bands are 207.34, 219.81, 219.10 for R, G and B bands, respectively. Also, the RMSE values obtained for the approximation sub-bands are higher as compared to the detail wavelet sub-bands. This is attributed to the range of values of wavelet coefficients in detail sub-bands when compared to the lower resolution image obtained in the approximation wavelet sub-bands.

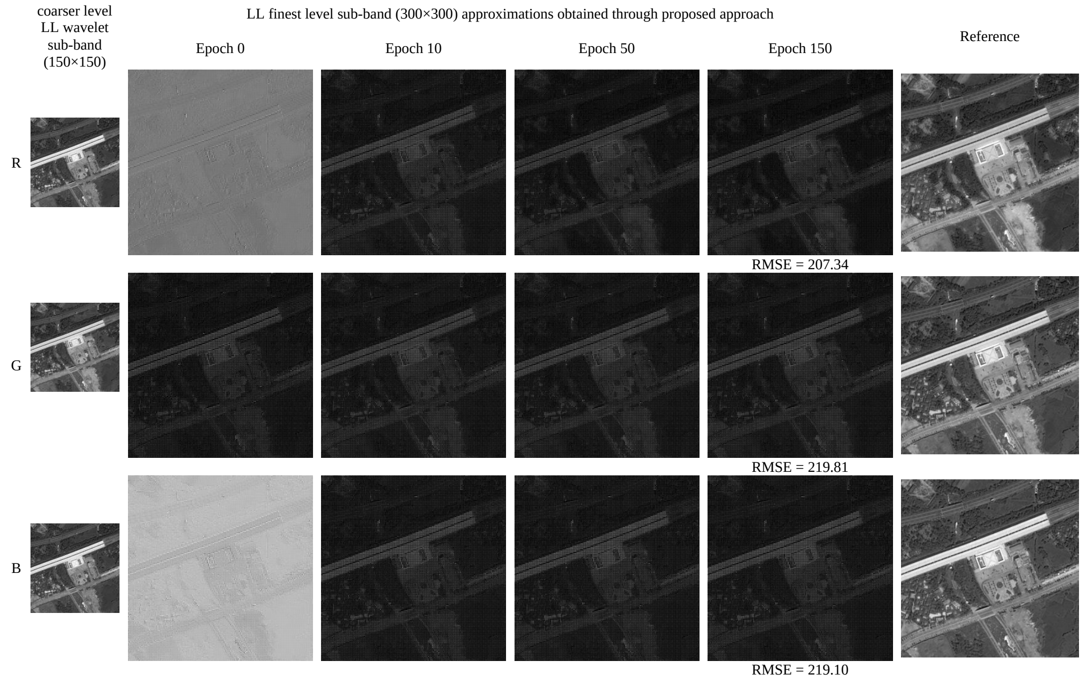

coarser level LL wavelet sub-band (150×150)
LL finest level sub-band (300×300) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 207.34
RMSE = 219.81
RMSE = 219.10

Figure 5.9: Qualitative results of sub-band approximations associated to LL wavelet sub-band of an image belonging to railway station category in the AID archive.

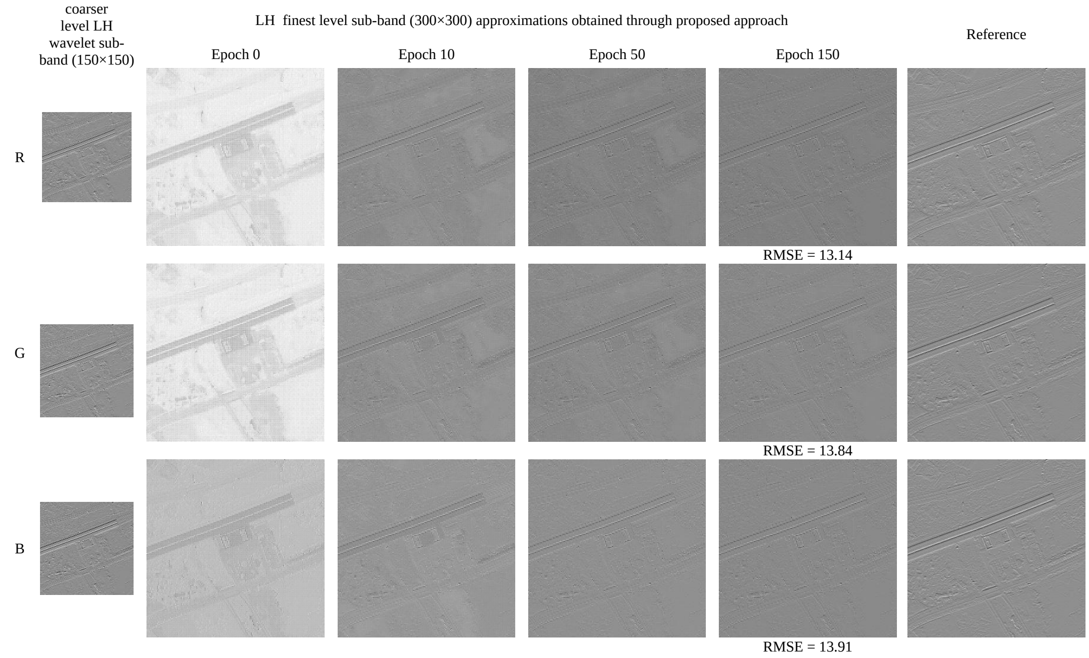

coarser level LH wavelet sub-band (150×150)
LH finest level sub-band (300×300) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 13.14
RMSE = 13.84
RMSE = 13.91

Figure 5.10: Qualitative results of sub-band approximations associated to LH wavelet sub-band of an image belonging to railway station category in the AID archive.

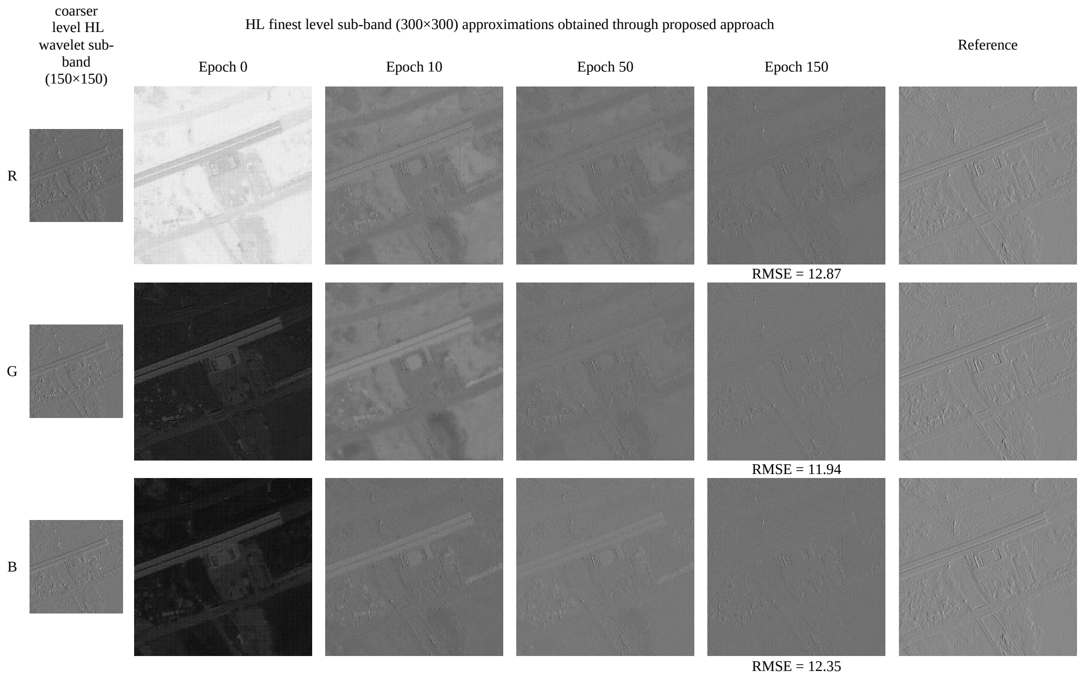

coarser level HL wavelet sub-band (150×150)
HL finest level sub-band (300×300) approximations obtained through proposed approach
Reference
Epoch 0
Epoch 10
Epoch 50
Epoch 150
R
G
B
RMSE = 12.87
RMSE = 11.94
RMSE = 12.35

Figure 5.11: Qualitative results of sub-band approximations associated to HL wavelet sub-band of an image belonging to railway station category in the AID archive.

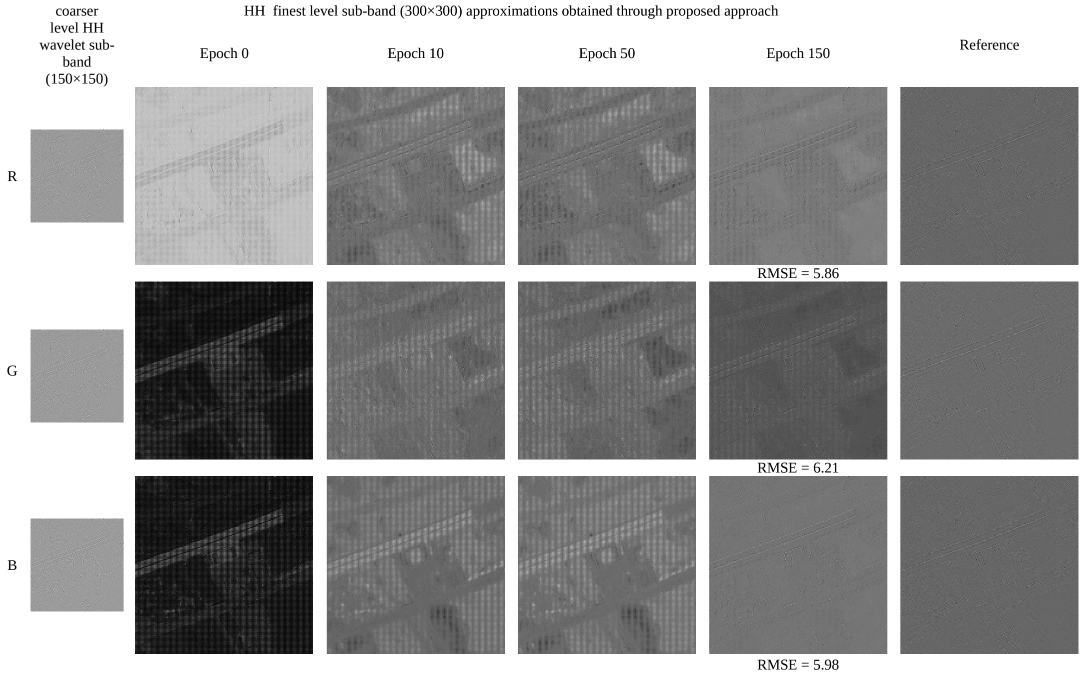

coarser level HH wavelet sub-band (150×150)
HH finest level sub-band (300×300) approximations obtained through proposed approach
Epoch 0
Epoch 10
Epoch 50
Epoch 150
Reference
R
G
B
RMSE = 5.86
RMSE = 6.21
RMSE = 5.98

Figure 5.12: Qualitative results of sub-band approximations associated to HH wavelet sub-band of an image belonging to railway station category in the AID archive.

Table 5.2: Classification accuracy and computational time for the proposed Approximation approach (NWPU-RESISC45 archive).

Proposed Approximation Approach		Accuracy (%)	Computational Time (sec)		
			Train	Validation	Test
Scenario 1	Approximating image (32 × 32) → (64 × 64) → (128 × 128) → (256 × 256)	73.27	8770.76	6.13	5.17
	Approximating finest level wavelet sub-bands **(32 × 32)→ (64 × 64)→ (128 × 128)**	**74.05**	**6739.87**	**5.28**	**5.68**
	Approximating finer level wavelet sub-bands (32 × 32) → (64 × 64)	65.42	568.99	0.38	0.51
Scenario 2	Approximating image (64 × 64) → (128 × 128) → (256 × 256)	80.09	8630.20	106.37	106.51
	Approximating finest level wavelet sub-bands **(64 × 64)→ (128 × 128)**	**79.92**	**8393.79**	**102.03**	**101.81**
	Approximating image (128 × 128) → (256 × 256)	78.54	8853.99	207.24	206.81

5.4.2 Results of the Proposed Approximation Approach for NWPU-RESISC45 and AID Archive

This subsection presents the classification accuracies and the computational time required by the proposed approach. For the following experiments, the feature extraction and classification steps of the proposed approach have been based on the AlexNet model. Table 5.2 reports the performance of the proposed approximation approach (both Scenario 1 and 2) for the NWPU-RESISC45 benchmark archive. Note that the computational time includes the decoding time required for the considered images. From the numbers in Table 5.2 associated to Scenario 1, one can notice that the proposed approach employs the coarsest level wavelet sub-bands (32 × 32) to approximate: i) the image (256 × 256) after applying three transposed convolutional layers; ii) the finest level wavelet sub-band (128 × 128) after applying two transposed convolutional layers; and iii) the finer level wavelet sub-band (64 × 64) after applying one transposed convolutional layer. As one can observe, approximating the finest level wavelet sub-band (128 × 128) achieves the best classification performance when compared to the other two cases. This is because when the coarsest level wavelet sub-bands are used to approximate image (which requires three transposed convolution layers), the details of the approximated fine-scaled objects are reduced. Nonetheless, when the finest level wavelet sub-bands (128 × 128) are approximated, (which requires only two transposed convolution layers) we gain in terms of both performance and computational time. Thus, we can conclude that, as the number of layers used for approximation decreases, the performance increases. In the third case, where the finer level sub-bands (64 × 64) is approximated (using one transposed convo-

Table 5.3: Classification accuracy and computational time for the proposed Approximation approach (AID archive).

Proposed Approximation Approach		Accuracy (%)	Computational Time (sec)		
			Train	Validation	Test
Scenario 1	Approximating image (75 × 75) → (150 × 150) → (300 × 300) → (600 × 600)	74.64	14539.87	14.26	14.03
	Approximating finest level wavelet sub-bands (75 × 75) → (150 × 150) → (300 × 300)	76.92	13598.14	13.87	14.91
	Approximating finer level wavelet sub-bands **(75 × 75)→(150 × 150)**	**77.34**	**10115.91**	**8.62**	**9.90**
Scenario 2	Approximating image (150 × 150) → (300 × 300) → (600 × 600)	79.91	14183.46	253.98	279.28
	Approximating finest level wavelet sub-bands **(150 × 150)→(300 × 300)**	**79.24**	**13847.33**	**224.36**	**227.34**
	Approximating image (300 × 300) → (600 × 600)	78.52	14964.64	326.34	331.65

lutional layers), the size of the approximated wavelet sub-bands does not provide enough image information to train the classifier. Thus, the resulting classification accuracy is the lowest (i.e. 65.42%) when compared to the other two cases.

From the numbers in Table 5.2 associated to Scenario 2, one can see that the proposed approach has used the decoded finer level wavelet sub-bands (64×64) to approximate: i) the image (256×256) after applying two transposed convolutional layers; and ii) the finest level wavelet sub-bands (128×128). In the third case, the proposed approach uses the decoded finest level wavelet sub-bands (128×128) to approximate the image (256×256). As one can see, all the three cases report almost similar classification accuracies with very small differences. However, if we compare the computational times, we can observe that the training time required for approximating the finest level wavelet sub-bands is lower than the time required to approximate the images. The training time required when the finest level wavelet sub-bands are used is 6739.87 sec. In the classification phase, the proposed approach takes 206.81 sec when the image is approximated after decoding two wavelet decomposition levels. In the first case, where the image is approximated using the the finer level wavelet sub-bands (64×64), the required computational time is only 106.51 sec. The overall gain is achieved when the finest level wavelet sub-bands (128×128) is approximated.

When we compare Scenarios 1 and 2 (Table 5.2), we can notice that the proposed approach attains good classification accuracies when the finest level wavelet sub-bands are used. If we analyze the performance of the proposed approach when the finest level wavelet sub-bands (128×128) are obtained, it achieves an accuracy of 74.04% when the coarsest level wavelet sub-bands are used with a required classification time as 5.68

sec. In the other case, approximating the finest level wavelet sub-bands (128 × 128) after decoding results in 79.92% classification accuracy with a higher computational time of 101.81 sec. We can observe that the proposed approach obtains accuracy of 74.05% when only the coarsest level wavelet sub-bands are used with a significantly reduced computational time 5.68 sec. If we perform one level wavelet decoding to obtain the finer level (64 × 64) wavelet sub-bands, which is used to approximate the finest level wavelet sub-bands (128 × 128), we notice an increase of 5.87% in classification accuracy. This shows that the proposed approach achieves reasonable classification performance with the coarsest wavelet sub-bands. Thus, the experimental results demonstrate that the finest level wavelet sub-bands (partially decoded domain) provide sufficient information for an efficient scene classification with reduced computational time.

Table 5.3 reports the performance of the proposed approach on the AID benchmark archive. From Table 5.3 (Scenario 1 and 2), the proposed approach employs the coarsest and the decoded finer level wavelet sub-bands (75 × 75) to approximate the finer level wavelet sub-bands (the image itself). While analyzing the part of Table 5.3 associated to Scenario 1, we can notice that the proposed approach employs the coarsest level wavelet sub-band to approximate: i) the image level information (600 × 600); ii) the finest level wavelet sub-bands (300 × 300); and iii) the finer level wavelet sub-band (150 × 150). From the results, we can observe a good accuracy of 77.34% when we approximate the finest level wavelet sub-band (150 × 150). This is because, the size of the coarsest level wavelet sub-band employed provides sufficient information to approximate finer level sub-bands (150 × 150) without large information loss with only one transposed convolutional layer. However, when we use two or more transposed convolutional layers to approximate finer level wavelet sub-bands (the image itself), the accuracy is reduced. This is because the coarsest level wavelet sub-bands (75 × 75) introduce checkerboard artifacts when two or more transposed transposed convolutional layers are included. In addition, the training and classification times required are 10115.91 sec and 9.90 sec, respectively. Thus, it requires minimum decoding, which reduced the additional overhead required before classification.

In the part of Table 5.3 associated to Scenario 2, the proposed approach employs: i) the finer level wavelet sub-bands (150 × 150) to approximate the image level information (600 × 600); ii) the finer level wavelet sub-bands (150 × 150) finest level wavelet sub-bands (300 × 300); and iii) the finest level wavelet sub-bands (300 × 300) to approximate the image (600 × 600). If we compare the classification accuracies, we can observe that the highest classification accuracy of 79.91% is achieved when the finer level wavelet sub-bands (150 × 150) are used to approximate the image (600 × 600). However, the training time required to approximate the finest level wavelet (300 × 300) from the finer level wavelet

Table 5.4: Classification accuracy and computational time for the proposed approximation approach and a standard CNN (NWPU-RESISC45 archive).

Model	Method		Accuracy (%)	Computational Time (sec)		
				Train	Validation	Test
AlexNet	Proposed Approximation Approach	Approximating finest level wavelet sub-bands (32 × 32) → (64 × 64) → (128 × 128)	74.05	6739.87	5.28	5.68
		Approximating finest level wavelet sub-bands (64 × 64) → (128 × 128)	79.92	8393.79	102.03	101.81
	Standard CNN	Fully decompressed image (256 × 256)	80.11	7478.89	305.13	306.24
		Without any decompression (32 × 32)	54.01	314.20	0.13	0.12
ResNet50	Proposed Approximation Approach	Approximating finer level wavelet sub-bands (75 × 75) → (150 × 150)	85.91	16953.31	15.61	16.01
		Approximating finest level wavelet sub-bands (150 × 150) → (300 × 300)	93.98	18992.71	124.32	125.64
	Standard CNN	Fully decompressed image (600 × 600)	94.85	17234.51	326.63	325.98
		Without any decompression (75 × 75)	76.31	763.24	3.64	3.98

sub-bands is lower when compared to the other two cases and the classification accuracy is 79.24% which is very close to the highest one. In addition, this last case has also the lowest classification time. From the experimental results, we can conclude that, if the size of the coarsest level wavelet sub-bands is large enough (e.g. as in the case of (75 × 75) AID archive), the proposed approach requires only one approximation level to achieve an acceptable classification accuracy.

5.4.3 Comparison of the Proposed Approach with a Standard CNN.

In this subsection, we compare the effectiveness of the proposed approach with: i) a standard-CNN model where full decompression of images is required; and ii) a standard-CNN model that takes as input the coarsest level wavelet sub-bands (which can be obtained from the codestreams of the compressed image). For the following experiments, the feature extraction and classification parts are based on the ResNet50 model. Tables 5.4 and 5.5 report the classification accuracies and computational times for the NWPU-RESISC45 and AID image archives, respectively. It is worth noting that during classification the proposed approach requires only the codestreams associated with the coarsest level wavelet sub-bands, whereas the standard-CNN model requires the fully decompressed images. By analyzing the tables one can observe that the computational time required by the proposed approach is significantly reduced when compared to that of the standard-CNN model. In addition, we can also notice that the proposed approach attains almost similar classification accuracies when compared to the standard-CNN model

Table 5.5: Classification accuracy and computational time for the proposed approximation approach and a standard CNN (AID archive).

Model	Method		Accuracy (%)	Computational Time (sec)		
				Train	Validation	Test
AlexNet	Proposed Approximation Approach	Approximating finer level wavelet sub-bands (75 × 75) → (150 × 150)	77.34	10115.91	8.62	9.90
		Approximating finest level wavelet sub-bands (150 × 150) → (300 × 300)	79.24	13847.33	224.36	227.34
	Standard CNN	Fully decompressed image (600 × 600)	79.54	12582.21	412.37	422.84
		Without any decompression (75 × 75)	61.91	946.23	7.90	8.25
ResNet50	Proposed Approximation Approach	Approximating finer level wavelet sub-bands (75 × 75) → (150 × 150)	84.92	17256.34	14.32	14.13
		Approximating finest level wavelet sub-bands (150 × 150) → (300 × 300)	92.24	b24356.75	298.26	299.50
	Standard CNN	Fully decompressed image (600 × 600)	93.01	21731.25	443.91	443.12
		Without any decompression (75 × 75)	69.78	1231.24	13.56	13.14

that uses fully decompressed images. On the contrary, if we perform classification using the coarsest level wavelet sub-bands, the classification accuracy is significantly reduced.

By analyzing the AlexNet model results for NWPU-RESISC45 archive (Table 5.4), we can notice that the classification accuracy obtained by using fully decompressed images with a standard CNN is 80.11%, with a classification time (i.e. test time) of 306.24 sec. The proposed approach results in a very similar classification accuracy of 79.92% when only one level of decoding is performed with a lower classification time of 101.81 sec. When the coarsest level wavelet sub-bands (32 × 32) are used to approximate finest level wavelet sub-bands (128 × 128), the required classification time is of more than an order of magnitude smaller at the cost of almost 5% lower classification accuracy. When the coarsest level wavelet-subbands are used in the standard CNN, we obtain the lowest classification accuracy with the lowest classification time. By analyzing the ResNet50 model results for the AID archive (Table 5.4), the classification accuracy obtained by fully decompressing the images is 94.85% with a classification time of 325.98 sec. The proposed approach results again in a very similar classification accuracy of 93.98% by reducing classification time (i.e. test time) to 125.64 sec.

By analyzing the AlexNet model results for NWPU-RESISC45 archive (Table 5.5), we observe that the proposed approach results in a classification accuracy of 77.34% when the coarsest level wavelet sub-bands are used, with a classification time of 9.90 sec. When we compare the performance of the proposed approach with the standard-CNN, although the classification accuracy is reduced by 2.20%, there is a significant gain in terms of the classification time that is reduced to 9.90 sec. Also, it is important to note that the proposed approach reaches a classification accuracy of 79.24% which is similar to that obtained by the standard-CNN approach that requires fully decompressed images. By analyzing ResNet50 model results for the AID archive (Table 5.5), the classification accuracy obtained by fully decompressing the images is 93.01% with a computational time of 444.12 sec. The proposed approach results in a similar classification accuracy of 92.24% with a computational time 299.50 sec. By analyzing the results, one can conclude that the proposed approach minimizes the computational time considerably when compared to the standard-CNN model. In addition, by using a powerful CNN model like ResNet50, the performance is also improved. However, this is achieved at the cost of increasing the computational time.

5.5 Conclusion

In this chapter, a novel approximation approach has been presented to perform RS image scene classification in the JPEG 2000 compressed domain by using DNNs. The proposed approach minimizes the amount of image decoding by training the DNN model using the approximations obtained from the codestreams associated to the coarser level wavelet sub-bands. To this end, the proposed method initially takes the codestreams associated to the coarsest level wavelet sub-bands to feed the model with a few transposed convolutional layers in order to approximate the finer level wavelet sub-bands. Then, the high-level semantic content of the approximated images is obtained through five convolutional layers followed by two FC layers. The aim of the transposed convolutional layers is to approximate the finer level wavelet sub-bands without requiring to decode the images (in order to obtain the features for scene classification). This significantly reduces the decoding time required for scene classification, which is the dominant aspect while performing scene classification in compressed RS image archives. Then, the features obtained from the finer level wavelet sub-bands are obtained through the convolutional layers. In addition to the classification loss, the estimation loss is also calculated between the approximated wavelet coefficients and the original wavelet coefficients. Accordingly, during training, the model learns the intrinsic behavior of the original compressed domain features that are reduced through the estimation loss. In addition, the time required to fully-decode the images is considerably minimized during the classification phase. The proposed model is then used to perform scene classification with JPEG 2000 partially compressed RS images.

Experimental results in terms of scene classification accuracy and computational gain on two benchmark archives demonstrates the effectiveness of the proposed approach. This is mainly related to the significant reduction of the decoding time associated with the use of a large amount of compressed images. As there is a trade-off between the computational gain and the classification accuracy based on the number of transposed convolutional layers, one can always choose the number of layers depending on the requirements in computational time and accuracy. The qualitative images obtained from the approximations show that the proposed approach operates efficiently only with the original coarsest level wavelet coefficients as input source. The results obtained from the experiments demonstrate the ability of the proposed approach:

1. To perform image scene classification within the compressed domain.
2. To significantly improve the computational gain by minimizing the amount of decompression required compared to the existing scene classification methods (which works in uncompressed domain).

Chapter 6

Conclusions and Future developments

In this thesis, we presented novel methods to perform CBIR and scene classification in real large-scale compressed RS image archives. Current trends in increase in the volume of compressed RS data demands the need to exploit the possibility to efficiently utilize the information obtained from compressed images in big data archives. Although the existing state-of-the-art presents several studies to achieve image retrieval and scene classification, they require fully decompressed images as input. Limited studies exploit the possibility to address the challenges and possibilities to perform scene classification and retrieval in compressed RS image archives. The proposed methods thereby contributes a valuable effort with a possibility of a new interesting research direction to solve big data problems in compressed RS image archives. The focus of the proposed thesis relies on the accurate image characterization within the compressed domain. This thesis proposed two novel contributions that showcased significant improvements in performance as well as computational time when compared to the state-of-the-art methods.

In the first contribution of the thesis, we proposed a novel RS CBIR system that achieves image characterization and retrieval in the JPEG 2000 compressed RS image archives. The proposed method significantly reduces the decoding time required by all the images in the archive by eliminating irrelevant images using hierarchically achieved partially decoded image descriptors. Experimental results obtained on two benchmark archives pointed out that the proposed approach resulted in similar retrieval performance when compared to the fully decompressed domain. In addition, the proposed method demonstrated a significant reduction in the decoding as well as the retrieval time in compressed RS image archives. The overall sharp improvement in performance (in terms of computational time) is achieved because of the elimination of irrelevant images in reference to the query image during the early retrieval stages that further reduces the decoding as well as required retrieval time. Moreover, the proposed approach can be adopted for archives that uses other compression algorithms by modifying the image description used in the proposed approach.

In the second contribution of the thesis, we proposed a novel approximation approach that is achieved in an end-to-end DNN architecture to achieve scene classification in compressed images. The proposed method showcase the potential of DNNs to achieve scene classification for compressed RS image archives with significantly reduced computational time when compared to the state-of-the-art systems. The results confirm the effectiveness of the proposed approximation approach (that uses the transposed convolutional layers) to efficiently characterize the compressed domain wavelet subband information. The number of transposed convolutional layers can be selected based on the user requirements as there is trade-off between the computational time and the classification performance. The proposed approach could be adopted by other compression algorithms by modifying the transposed convolutional layers. In view of the growth of RS big data archives, the second contribution introduces a research direction very important for operating scene classification directly on compressed archives. Note that the proposed approach is not limited to JPEG 2000 compressed archives but can be directly applied to any image archive that considers wavelet based compression approach. In addition, it can be adapted to be used in the framework of other compression algorithms by modifying the technique used for approximating the compressed domain features.

In the thesis, we explored the potential of developing image retrieval and scene classification approaches for compressed RS image archives. By analysing the experimental results obtained from the proposed methods, we observe a few interesting research directions as part of the future developments of the presented methods. First, we plan to validate the proposed progressive image retrieval system on larger image archives. Further, time will be devoted to develop methods that could combine the features obtained from different spectral bands to speed up the retrieval rate. In view of the second contribution of the thesis, we plan to explore scene classification in the context of Generative Adversarial Networks in compressed domain. The ability of Generative Adversarial Networks to produce images can be adapted to approximate the wavelet subband information which can be utilised to address scene classification problems. Finally, we aim to focus on developing novel methods to perform scene classification and retrieval when the image compression is achieved within the DNNs. Recent advances in deep learning has demonstrated its ability to compress RS images. Several novel models that considers RNNs, LSTMs, GANs has been proposed to compress the data. In view of this, we plan to study the development of models that can extract features within a deeply compressed domain.

Chapter 7

List of Publications

International Journal Publications

1. Byju, A. P., Demir, B., Bruzzone, L., **’A Progressive Content Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives’**, IEEE Transactions on Geoscience and Remote Sensing, accepted for publication, 2020.

2. Byju, A. P., Sumbul, G., Demir, B., Bruzzone, L., **’Remote Sensing Image Scene Classification with Deep Neural Networks in JPEG 2000 Compressed Domain’**, IEEE Transactions on Geoscience and Remote Sensing, accepted for publication, 2020.

Conferences

1. Byju, A. P., Sumbul, G., Demir, B., Bruzzone, L., **’Approximating JPEG 2000 wavelet representation through deep neural networks for remote sensing image scene classification.’**, SPIE 2019, SPIE Conference on Image and Signal Processing for Remote Sensing XXV, Strasbourg, France, 10-14 September 2019. (oral presentation)

2. Byju, A. P., Demir, B., Bruzzone, L., **’A novel coarse-to-fine remote sensing image retrieval system in JPEG-2000 compressed domain”, SPIE Image and Signal Processing for Remote Sensing.’**, SPIE 2018, SPIE Conference on Image and Signal Processing for Remote Sensing XXIV, Berlin, Germany, 10-14 September 2018. (oral presentation)

Other Journal Publications

1. Byju, A. P., Kumar, A., Stein, A., Kumar, A. S., **’Combining the FCM Classifier with Various Kernels to Handle Non-linearity of Class Boundaries’**, Journal of the Indian Society of Remote Sensing, 46(9), 1519-1526, 2018.

Bibliography

[1] N. Elgendy and A. Elragal, “Big Data Analytics: A Literature Review Paper,” in *Industrial Conference on Data Mining*, pp. 214–227, 2014.

[2] P. Kempeneers and P. Soille, “Optimizing Sentinel-2 image selection in a Big Data Context,” *Big Earth Data*, vol. 1, no. 1-2, pp. 145–158, 2017.

[3] T. Hahmann, “10 000 000 Gigabyte Sentinel am EOC.”

[4] “Copernicus: Sentinel-2 The Optical Imaging Mission for Land Services.”

[5] C. Galeazzi, A. Sacchetti, A. Cisbani, and G. Babini, “The Prisma Program,” in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 105–108, 2008.

[6] P. L. Dragotti, G. Poggi, and A. R. Ragozini, “Compression of Multispectral Images by Three-Dimensional SPIHT Algorithm,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 416–428, 2000.

[7] S. Zhou, C. Deng, B. Zhao, Y. Xia, Q. Li, and Z. Chen, “Remote Sensing Image Compression : A Review,” in *IEEE International Conference on Multimedia Big Data*, pp. 406–410, 2015.

[8] A. Zabala, X. Pons, R. Díaz-Delgado, F. García, F. Aulí-Llinàs, and J. Serra-Sagristà, “Effects of JPEG and JPEG2000 Lossy Compression on Remote Sensing Image Classification for Mapping Crops and Forest Areas,” in *IEEE International Symposium on Geoscience and Remote Sensing*, pp. 790–793, 2006.

[9] C. Lambert-Nebout, C. Latry, G. A. Moury, C. Parisot, M. Antonini, and M. Barlaud, “On-board Optical Image Compression for Future High-Resolution Remote Sensing Systems,” in *Applications of Digital Image Processing XXIII International Society for Optics and Photonics*, vol. 4115, pp. 332–346, 2000.

[10] A. Zabala, J. Gonzalez-Conejero, J. Serra-Sagrist‘, and X. Pons, “JPEG 2000 Encoding of Images with NODATA Regions for Remote Sensing Applications,” *Journal of Applied Remote Sensing*, vol. 4, no. 1, p. 041793, 2010.

[11] N. D. Memon, K. Sayood, and S. S. Magliveras, “Lossless Compression of Multispectral Image Data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 2, pp. 282–289, 1994.

[12] A. E. Ali and D. A. Algarni, “The Effect of Data Compression on Remote Sensing Image Classification,” *Journal of King Saud University - Engineering Sciences*, vol. 12, no. 2, pp. 187–196, 2000.

[13] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. G. Marchetti, and S. D’Elia, “Information Mining in Remote Sensing Image Archives: System Concepts,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2923–2937, 2003.

[14] M. Cagnazzo, G. Poggi, and L. Verdoliva, "Region-Based Transform Coding of Multispectral Images," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2916–2926, 2007.

[15] G. Gelli and G. Poggi, "Compression of Multispectral Images by Spectral Classification and Transform Coding," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 476–489, 1999.

[16] J. Mielikainen and P. Toivanen, "Clustered DPCM for the Lossless Compression of Hyperspectral Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2943–2946, 2003.

[17] G. K. Wallace, "The JPEG Still Picture Compression Standard," *IEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii – xxxiv, 1991.

[18] D. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice.* Springer Publishing Company, 2013.

[19] "Sentenel-2 User Handbook," tech. rep., 2013.

[20] "PRISMA (Prototype Research Instruments and Space Mission technology Advancement)."

[21] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2018.

[22] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Region-Based Retrieval of Remote Sensing Images Using an Unsupervised Graph-Theoretic Approach," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 987–991, 2016.

[23] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep Learning in Remote Sensing Applications: A Meta-Analysis And Review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, no. April, pp. 166–177, 2019.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* 2016.

[25] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[26] X. Zheng, Y. Yuan, and X. Lu, "A Deep Scene Representation for Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4799–4809, 2019.

[27] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote Sensing Image Retrieval Using Convolutional Neural Network Features And Weighted Distance," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 10, pp. 1535–1539, 2018.

[28] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[29] L. Bashmal, Y. Bazi, H. AlHichri, M. M. AlRahhal, N. Ammour, and N. Alajlan, "Siamese-GAN: Learning Invariant Representations for Aerial Vehicle Image Categorization," *Remote Sensing*, vol. 10, no. 2, pp. 1–19, 2018.

[30] Y. Duan, X. Tao, M. Xu, C. Han, and J. Lu, “GAN-NL: Unsupervised Representation Learning for Remote Sensing Image Classification,” in *iIEE Global Conference on Signal and Information Processing, GlobalSIP*, pp. 375–379, 2018.

[31] L. Mou, P. Ghamisi, and X. X. Zhu, “Deep Recurrent Neural Networks For Hyperspectral Image Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[32] J. Zhang, W. Geng, X. Liang, J. Li, L. Zhuo, and Q. Zhou, “Hyperspectral Remote Sensing Image Retrieval System Using Spectral and Texture Features,” *Applied Optics*, vol. 56, no. 16, pp. 4785–4796, 2017.

[33] B. Demir, M. Ieee, and L. Bruzzone, “A Novel Active Learning Method in Relevance Feedback for Content Based Remote Sensing Image Retrieval,” *IEEE GeoScience and Remote Sensing*, vol. 53, no. 5, pp. 2323 – 2334, 2014.

[34] T. Reato, B. Demir, and L. Bruzzone, “An Unsupervised Multicode Hashing Method for Accurate and Scalable Remote Sensing Image Retrieval,” *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 276–280, 2019.

[35] W. Song, S. Li, and J. A. Benediktsson, “Deep Hashing Learning for Visual and Semantic Retrieval of Remote Sensing Images,” *arXiv preprint*, pp. 1–11, 2019.

[36] Y. Yang and S. Newsam, “Geographic Image Retrieval Using Local Invariant Features,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818–832, 2013.

[37] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, and Y. Zhong, “AID : A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification,” *IEEE Transactions on GeoScience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[38] Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, and D. Li, “Global-Local Attention Network for Aerial Scene Classification,” *IEEE Access*, pp. 67200–67212, 2019.

[39] X. Han, Y. Zhong, B. Zhao, and L. Zhang, “Scene Classification based on a Hierarchical Convolutional Sparse Auto-encoder for High Spatial Resolution Imagery,” *International Journal of Remote Sensing*, vol. 38, no. 2, pp. 514–536, 2017.

[40] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, “Using Convolutional Features and a Sparse Autoencoder for Land-Use Scene Classification,” *International Journal of Remote Sensing*, vol. 37, no. 10, pp. 2149–2167, 2016.

[41] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene Classification with Recurrent Attention of VHR Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2019.

[42] X. Pan, J. Zhao, and J. Xu, “A Scene Images Diversity Improvement Generative Adversarial Network for Remote Sensing Image Scene Classification,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019.

[43] G. Cheng, J. Han, and X. Lu, “Remote Sensing Image Scene Classification: Benchmark and State of the Art,” *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[44] Satellite Imaging Corporation, “Spot Satellite Technical Data,” tech. rep., 2003.

[45] A. Kumar, R. Kumaran, S. Paul, and R. M. Parmar, “Low Complex ADPCM Image Compression Technique,” in *Computational Intelligence and Information Technology*, pp. 318–321, 2013.

[46] M. Rabbani and R. Joshi, “An overview of the JPEG 2000 still image compression standard,” *Signal Processing: Image Communication*, vol. 17, pp. 3–48, 2002.

[47] A. Skodras, C. Christopoulos, and T. Ebrahimi, “The JPEG 2000 Still Image Compression Standard,” *IEEE Signal Processing Magazine*, pp. 36–58, 2001.

[48] S. Mallat and S. Zhong, “Characterization of Signals from Multiscale Edges,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, 1992.

[49] A. Gersho, “Quantization,” *IEEE communications society magazine*, vol. 15, no. 5, pp. 16–29, 1977.

[50] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.

[51] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, “ChangeNet: A Deep Learning Architecture for Visual Change Detection,” in *European Conference on Computer Vision*, 2018.

[52] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation Functions: Comparison of Trends in Practice and Research for Deep Learning,” in *arXiv preprint*, pp. 1–20, 2018.

[53] A. Teynor, M. Wolfgang, and K. Wolfgang, “Compressed Domain Image Retrieval Using JPEG2000 and Gaussian Mixture Models,” in *International Conference on Advances in Visual Information Systems*, pp. 132–142, 2005.

[54] C. A. Waring and X. Liu, “Face Detection using Spectral Histograms and SVMs,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 3, pp. 467–476, 2005.

[55] O. Regniers, J. P. Da Costa, G. Grenier, C. Germain, and L. Bombrun, “Texture Based Image Retrieval and Classification of Very High Resolution Maritime Pine Forest Images,” in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4038–4041, 2013.

[56] S. D. Newsam and C. Kamath, “Retrieval Using Texture Features in High-Resolution Multispectral Satellite Imagery,” *Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI*, vol. 5433, p. 21, 2004.

[57] R. M. Haralick, K. Shanmugham, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transactions on Systems, Man and Cyberbectics*, no. 6, pp. 610–621, 1973.

[58] I. Tekeste and B. Demir, “Advanced Local Binary Patterns for Remote Sensing Image Retrieval,” in *International Conference on Geoscience and Remote Sensing Symposium*, pp. 6855–6858, 2018.

[59] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[60] S. Liao, M. W. Law, and A. C. Chung, “Dominant Local Binary Patterns for Texture Classification,” *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 1107–1118, 2009.

[61] Z. Guo, L. Zhang, and D. Zhang, “Rotation Invariant Texture Classification using LBP Variance (LBPV) with Global Matching,” *Pattern Recognition*, vol. 43, no. 3, pp. 706–719, 2010.

[62] Z. Shao, W. Zhou, L. Zhang, and J. Hou, “Improved Color Texture Descriptors for Remote Sensing Image Retrieval,” *Journal of Applied Remote Sensing*, vol. 8, no. 1, p. 083584, 2014.

[63] E. Aptoula, “Remote Sensing Image Retrieval With Global Morphological Texture Descriptors,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 3023–3034, 2014.

[64] S. Valero, J. Chanussot, J. A. Benediktsson, H. Talbot, and B. Waske, “Advanced Directional Mathematical Morphology for the Detection of the Road Network in Very High Resolution Remote Sensing Images,” *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1120–1127, 2010.

[65] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[66] B. Demir and L. Bruzzone, “Hashing-Based Scalable Remote Sensing Image Search and Retrieval in Large Archives,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 2, pp. 892–904, 2016.

[67] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating Local Descriptors into a Compact Image Representation,” in *IEEE conference on Computer Vision and Pattern Recognition Pattern Recognition*, pp. 3304–3311, 2010.

[68] P. Napoletano, “Visual Descriptors for Content-based Retrieval of Remote- Sensing Images,” *International Journal of Remote Sensing*, vol. 39, no. 5, pp. 1343–1376, 2018.

[69] W. Zhou, S. Newsam, C. Li, and Z. Shao, “Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval,” *Remote Sensing*, vol. 9, no. 5, 2017.

[70] S. Roy, E. Sangineto, and N. Sebe, “Deep Metric and Hash-Code Learning for Content-Based Image Retrieval of Remote Sensing,” in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 4543–4546, 2018.

[71] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, “Fusion of Multiscale Convolutional Neural Networks for Building Extraction in Very High-Resolution Images,” *Remote Sensing*, vol. 11, no. 3, pp. 1–17, 2019.

[72] Q. Bao and P. Guo, “Comparative studies on similarity measures for remote sensing image retrieval,” in *IEEE International Conference on Systems, Man and Cybernetics*, pp. 1112–1116, 2004.

[73] C. Ma, Q. Dai, J. Liu, S. Liu, and J. Yang, “An improved SVM model for relevance feedback in remote sensing image retrieval,” *International Journal of Digital Earth*, vol. 7, no. 9, pp. 725–745, 2014.

[74] R. Liu, Y. Wang, T. Baba, D. Masumoto, and S. Nagata, “SVM-based active feedback in image retrieval using clustering and unlabeled data,” *Pattern Recognition*, vol. 41, no. 8, pp. 2645–2655, 2008.

[75] B. Demir and L. Bruzzo, “Kernel-Based Hashing for Content-Based Image Retrieval in Large Remote Sensing Data Archive,” in *International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3542–3545, 2014.

[76] D. Ye, Y. Li, C. Tao, X. Xie, and X. Wang, “Multiple Feature Hashing Learning for Large-scale Remote Sensing Image Retrieval,” *ISPRS International Journal of Geo-Information*, vol. 6, no. 11, 2017.

[77] B. D. Liu, J. Meng, W. Y. Xie, S. Shao, Y. Li, and Y. Wang, “Weighted Spatial Pyramid Matching Collaborative Representation for Remote-Sensing-Image Scene Classification,” *Remote Sensing*, vol. 11, no. 5, pp. 1–18, 2019.

[78] J. Fan, H. L. Tan, M. Toomik, and S. Lu, "Spectral-Spatial Hyperspectral Image Classification Using Super-Pixel-Based Spatial Pyramid Representation," in *Image and Signal Processing for Remote Sensing XXII*, vol. 10004, p. 100040W, 2016.

[79] J. Deng, R. Socher, L.-J. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[80] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, G. Anguelov, D. Erhan, V. Vanhoucke, and Rabin, "Going Deeper with Convolutions," in *IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[81] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating Very Deep Convolutional Networks for Classification and Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.

[82] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe : An Open Source Convolutional Architecture for Fast Feature Embedding," in *ACM International Conference on Multimedia*, pp. 675–678, 2014.

[83] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating Rich Hierarchical Features for Scene Classification in Remote Sensing Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 9, pp. 4104–4115, 2017.

[84] R. Gaetano, D. Ienco, K. Ose, and R. Cresson, "A Two-Branch CNN Architecture for Land Cover Classification of PAN and MS Imagery," *Remote Sensing*, vol. 10, no. 11, 2018.

[85] Y. Boualleg, M. Farah, and I. R. Farah, "Remote Sensing Scene Classification Using Convolutional Features and Deep Forest Classifier," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, pp. 1944 – 1948, 2019.

[86] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote Sensing Image Scene Classification Using Rearranged Local Features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1779–1792, 2019.

[87] X. Ma, H. Wang, and J. Geng, "Spectral-Spatial Classification of Hyperspectral Image Based on Deep Auto-Encoder," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4073–4085, 2016.

[88] S. Hao, W. Wang, Y. Ye, T. Nie, and L. Bruzzone, "Two-Stream Deep Architecture for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2349–2361, 2018.

[89] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, and S. Marshall, "Novel Segmented Stacked Autoencoder for Effective Dimensionality Reduction and Feature Extraction in Hyperspectral Imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.

[90] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese Convolutional Neural Networks for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1200–1204, 2019.

[91] D. Ma, P. Tang, and L. Zhao, "SiftingGAN: Generating and Sifting Labeled Samples to Improve the Remote Sensing Image Scene Classification Baseline In Vitro," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1046–1050, 2019.

[92] S. Xu, X. Mu, D. Chai, and X. Zhang, "Remote Sensing Image Scene Classification based on Generative Adversarial Networks," *Remote Sensing Letters*, vol. 9, no. 7, pp. 617–626, 2018.

[93] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2092–2096, 2017.

[94] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative Adversarial Networks and Probabilistic Graph Models for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 8191–8192, 2018.

[95] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised Hyperspectral Image Classification Based on Generative Adversarial Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 212–216, 2018.

[96] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of Hyperspectral Images Based on Multiclass Spatial-Spectral Generative Adversarial Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5329–5343, 2019.

[97] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *International Conference on Learning Representations, ICLR*, pp. 1–16, 2016.

[98] M. I. Lakhal, H. Çevikalp, S. Escalera, and F. Ofli, "Recurrent Neural Networks for Remote Sensing Image Classification," *IET Computer Vision*, vol. 12, no. 7, pp. 1040–1045, 2018.

[99] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded Recurrent Neural Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5384–5394, 2019.

[100] E. Ndikumana, D. H. T. Minh, N. Baghdadi, D. Courault, and L. Hossard, "Deep Recurrent Neural Network For Agricultural Classification Using Multitemporal SAR Sentinel-1 for Camargue, France," *Remote Sensing*, vol. 10, no. 8, pp. 1–16, 2018.

[101] Z. Sun, L. Di, and H. Fang, "Using Long Short-Term Memory Recurrent Neural Network in Land Cover Classification on Landsat and Cropland Data Layer Time Series," *International Journal of Remote Sensing*, vol. 40, no. 2, pp. 593–614, 2019.

[102] Y. Hua, L. Mou, and X. X. Zhu, "Multi-label Aerial Image Classification Using a Bidirectional Class-Wise Attention Network," in *Joint Urban Remote Sensing Event*, pp. 1–4, 2019.

[103] G. Sumbul and B. Demir, "A Novel Multi-Attention Driven System for Multi-Label Remote Sensing Image Classification," in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 5726–5729, 2019.

[104] A. Descampe, C. De Vleeschouwer, P. Vandergheynst, and B. Macq, "Scalable Feature Extraction for Coarse-to-Fine JPEG 2000 Image Classification," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2636–2649, 2011.

[105] F. Zargari, A. Mosleh, and M. Ghanbari, "A Fast And Efficient Compressed Domain JPEG 2000 Image Retrieval Method," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1886–1893, 2008.

[106] F. Zargari, A. Mosleh, and M. Ghanbari, "Compressed Domain JPEG 2000 Image Indexing Method Employing Full Packet Header Information," in *International Workshop on Content-Based Multimedia Indexing*, pp. 410–416, 2008.

[107] M. A. Muqeet and R. S. Holambe, "Local Binary Patterns Based on Directional Wavelet Transform for Expression and Pose-Invariant Face Recognition," *Applied Computing and Informatics*, vol. 15, no. 2, pp. 163–171, 2019.

[108] T. Cevik, A. M. A. Alshaykha, and N. Cevik, "Performance Analysis of GLCM-based Classification on Wavelet Transform-Compressed Fingerprint Images," in *International Conference on Digital Information and Communication Technology and Its Applications, DICTAP 2016*, pp. 131–135, 2016.

[109] B. Attallah, A. Serir, Y. Chahir, and A. Boudjelal, "Histogram of Gradient and Binarized Statistical Image Features of Wavelet Subband-Based Palmprint Features Extraction," *Journal of Electronic Imaging*, vol. 26, no. 06, p. 1, 2017.

[110] Y. Dong and J. Ma, "Wavelet-Based Image Texture Classification Using Local Energy Histograms," *IEEE Signal Processing Letters*, vol. 18, no. 4, pp. 247–250, 2011.

[111] M. S. Allili, "Wavelet Modeling using Finite Mixtures of Generalized Gaussian Distributions: Application to Texture Discrimination and Retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1452–1464, 2012.

[112] M. N. Do and M. Vetterli, "Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[113] E. De Ves, D. Acevedo, A. Ruedin, and X. Benavent, "A Statistical Model for Magnitudes and Angles of Wavelet Frame Coefficients and its Application to Texture Retrieval," *Pattern Recognition*, vol. 47, no. 9, pp. 2925–2939, 2014.

[114] E. De Ves, D. Acevedo, A. Ruedin, and X. Benavent, "A Statistical Model for Magnitudes and Angles of Wavelet Frame Coefficients and its Application to Texture Retrieval," *Pattern Recognition*, vol. 47, no. 9, pp. 2925–2939, 2014.

[115] A. Barla, F. Odone, and A. Verri, "Histogram Intersection Kernel for Image Classification," in *International Conference on Image Processing*, vol. 3, (Barcelona, Spain), pp. III–513, 2003.

[116] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Efficient Learning with Sets of Features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[117] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2018.

[118] M. H. Pi, C. S. Tong, S. K. Choy, and H. Zhang, "A Fast and Effective Model for Wavelet Subband Histograms and Its Application in Texture Image Retrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3078–3088, 2006.

[119] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018.

[120] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[121] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.