Classification Using Sparse Representation and Applications to Skin Lesions Diagnosis

Long H. Ngo

THÈSE

pour obtenir le grade de

Docteur de l'Université Sorbonne Paris Nord

Discipline : "Signaux et Images"

Long H. NGO

Classification Using Sparse Representation and Applications to Skin Lesions Diagnosis

Directeur de thèse : **Pr. Emmanuel Viennet**

Co-encadrant de thèse : **Dr. Marie Luong**

Collaborateurs : **Pr. Nikolay M. Sirakov, Pr. Thuong Le-Tien**

JURY

Rapporteurs :	**M. Mouloud ADEL**	- Prof., Aix-Marseille Université, France
	M. Patrick SIARRY	- Prof., Université Paris-Est Créteil, France
Examinateurs :	**M. Thuong LE-TIEN**	- Prof., Hochiminh City Univ. of Technology, Vietnam
	Mme. Su RUAN	- Prof., Université de Rouen, France
	M. Nikolay M. SIRAKOV	- Prof., Texas A&M University-Commerce, USA
Co-encadrant :	**Mme. Marie LUONG**	- MCF, Université Sorbonne Paris Nord, France
Directeur :	**M. Emmanuel VIENNET**	- Prof., Université Sorbonne Paris Nord, France

Thesis

Submitted for the degree of Doctor

Université Sorbonne Paris Nord

Specialization : "Signal and Image Processing"

Long H. NGO

Classification Using Sparse Representation and Applications to Skin Lesions Diagnosis

Supervisor : **Pr. Emmanuel Viennet**

Co-supervisor : **Dr. Marie Luong**

Collaborators : **Pr. Nikolay M. Sirakov, Pr. Thuong Le-Tien**

Committee

Reviewers:	**Mouloud ADEL**	- Prof., Aix-Marseille Université, France
	Patrick SIARRY	- Prof., Université Paris-Est Créteil, France
Examiners:	**Thuong LE-TIEN**	- Prof., Hochiminh City Univ. of Technology, Vietnam
	Su RUAN	- Prof., Université de Rouen, France
	Nikolay M. SIRAKOV	- Prof., Texas A&M University-Commerce, USA
Co-Supervisor:	**Marie LUONG**	- MCF, Université Sorbonne Paris Nord, France
Supervisor:	**Emmanuel VIENNET**	- Prof., Université Sorbonne Paris Nord, France

L2TI – EA3043

École doctorale Galilée – Université Sorbonne Paris Nord

99, avenue Jean-Baptiste Clément

93430 Villetaneuse.

Résumé

La classification d'images est une discipline majeure en traitement d'images et en intelligence artificielle. La classification est d'une importance fondamentale pour qu'un système intelligent puisse exploiter et gérer efficacement l'information visuelle. L'objectif est de développer des algorithmes qui trouvent automatiquement la catégorie à laquelle appartient un échantillon d'image, à partir d'échantillons d'entraînement. Dans nos études, nous nous concentrons sur l'étude et le développement des algorithmes basés sur la représentation parcimonieuse pour la classification d'images, y compris, mais sans s'y limiter, les visages, les objets et les lésions cutanées. Cette étude met l'accent sur le développement des problèmes de classification basés sur la représentation parcimonieuse dans les domaines spécifiques tels que le domaine des ondelettes ou le domaine des ondelettes quaternioniques dans le but d'améliorer les performances de séparation des classes.

En outre, notre objectif est de mettre en œuvre une nouvelle méthode pour le diagnostic du mélanome assisté par ordinateur, réalisé à partir d'images dermoscopiques. Le mélanome est le type de cancer de la peau le plus mortel. Heureusement, les lésions cutanées sont curables si elles sont diagnostiquées et traitées suffisamment tôt. Pour cette raison, le diagnostic automatique du mélanome assisté par ordinateur suscite aujourd'hui un grand intérêt de la part des chercheurs.

Dans la première partie de cette étude, nous proposons une nouvelle méthode basée sur la représentation parcimonieuse, à savoir la classification basée sur la représentation parcimonieuse dans le domaine des ondelettes (SRWC), qui résout le problème du codage parcimonieux dans le domaine des ondelettes. Le cadre de la SRWC montre que les caractéristiques obtenues à partir de la transformation en ondelettes peuvent contribuer au processus de classification. En particulier, nous fusionnons les caractéristiques de l'image décrites par les informations complémentaires des coefficients d'ondelettes à basse fréquence et la représentation parcimonieuse pour améliorer les performances

de classification et en comparant avec des méthodes conventionnelles de classification par représentation parcimonieuse. Comme les ondelettes favorisent la parcimonie et fournissent des informations structurelles sur l'image, la méthode proposée augmente la précision de la classification. En outre, notre méthode peut naturellement gérer l'occlusion et la corruption des images.

Dans la deuxième partie de cette étude, nous étendons la méthode SRWC à l'espace 4D des quaternions pour développer une nouvelle méthode de classification basée sur la représentation parcimonieuse dans le domaine des ondelttes quaternioniques, appelé SRCQW (Sparse Representation based Classification in the Quaternion Wavelet domain). En particulier, cette méthode exploite la transformée en ondelettes quaternioniques, qui utilise les filtres et la transformée de Hilbert, pour générer les coefficients d'ondelettes quaternioniques. Comme pour la méthode précédente, nous n'utilisons que les caractéristiques quaternioniques décrites par les coefficients des sous-bandes de basse fréquence pour mapper le dictionnaire parcimonieux et le problème de classification dans l'espace quaternionique 4D. Pour calculer le vecteur quaternionique parcimonieux, nous formulons le modèle QWLasso (quaternion wavelet least absolute shrinkage and selection operator) en utilisant la minimisation du l_1 quaternionique. Pour résoudre le problème QWLasso, nous développons le nouvel algorithme QFISTA (quaternion fast iterative shrinkage-thresholding algorithm). La combinaison des ondelettes quaternionques, qui favorisent la parcimonie, et du modèle de représentation parcimonieuse garantit la convergence de la méthode proposée vers une grande précision de la classification.

Dans la troisième partie de l'étude, nous combinons le SRWC et le réseau de neurones (NN) pour pallier aux inconvénients des deux approches. Plus précisément, il s'agit d'une méthode de classification par apprentissage qui se base sur un modèle d'autoencodage convolutif (CAE), et sur une représentation parcimonieuse dans le domaine des ondelettes afin de classer les images étiquetées. Pour cela, nous appelons cette méthode CAE-SRWC. Ce travail est réalisé dans le cadre d'une collaboration avec un étudiant de Master. Dans l'approche proposée, la CAE apprend, avec une couche latente parcimonieuse, les codes parcimonieux des caractéristiques des ondelettes. Ensuite, un critère probabiliste basé sur les résidus est utilisé pour attribuer des étiquettes aux échantillons de test en fonction des codes parcimonieux estimés. En outre, la méthode proposée montre explicitement une réduction substantielle du nombre de paramètres du réseau par rapport aux méthodes récentes de réseaux de neurones.

L'efficacité des avancées théoriques et des méthodes proposées est validée expérimentalement en les appliquant à des bases de données couramment utilisées, telles que celles de visages et des objets, et en comparant leurs résultats avec celles des méthodes de pointe dans le domaine, y compris les méthodes de réseaux de neurones.

Dans la dernière partie du travail, nous démontrons les capacités des algorithmes proposés, notamment SRWC, SRCQW et CAE-SRWC, pour le traitement des images biomédicales en les appliquant à la classification des images de lésions cutanées. Les résultats obtenus montrent le potentiel des méthodes nouvellement développées, pour classer les images de lésions cutanées dermoscopiques. De plus, les trois approches proposées montrent leur supériorité pour la reconnaissance des images de mélanome avec de bons résultats de sensibilité.

Mots clés : Classification, Représentation parcimonieuse, Transformée en ondelettes discrète, Algèbre de quaternions, Transformée en ondelettes quaternionique, Lésions cutanées.

Abstract

Image classification, a key research in image processing and artificial intelligence, is of fundamental importance for an intelligent system to exploit and manage efficiently the visual information. The objective is to develop algorithms that automatically find the category, to which an image sample belongs, given training samples. In our studies, we focus on the research and applications of sparse representation based algorithms for image classification including but not limited to faces, objects and skin lesions. A key emphasis of this study is to formulate the sparse representation-based classification problems in specific domains, like wavelet and quaternion wavelet, in order to enhance classes separation performance.

Further, our goal is to implement the novel method to computer-assisted melanoma diagnosing, performed on dermoscopic images. Melanoma is the most deadly type of skin cancer. Fortunately, skin lesions are curable if they are diagnosed and treated early enough. Due to this reason, the automated computer-assisted melanoma diagnosing has attracted great interest to researchers nowadays.

In the first stage of the present study we propose a novel sparse representation based methods, namely Sparse Representation Wavelet based Classification (SRWC), solving the sparse coding problem in the wavelet domain. The SRWC framework shows that features obtained from wavelet transform can contribute to the classification process. In particularly, we fuse the image features described by the complementary information from the low-frequency wavelet coefficients and sparse representation to outperform the conventional sparse representation-based methods according to accuracy. As the wavelets promote sparsity and provide structural information about the image, the proposed method increases the accuracy of classification. Furthermore, our method can naturally handle occlusion and corruption in images.

In the second stage of the present study, we extend the SRWC method to the 4D space of quaternions to develop a novel method called Sparse Representation based

Classification in the Quaternion Wavelet domain (SRCQW). In particularly, this method exploits the quaternion wavelet transform, which considers the low, high-pass filters and their Hilbert transform calculated counterparts, to generate the quaternionic wavelet coefficients. Analogous to our previous work, we only use the quaternion features described by the coefficients from the low-frequency wavelet sub-bands to map the sparse dictionary and the classification problem onto the 4D quaternion space. To calculate the quaternion sparse vector, we formulate the quaternion wavelet least absolute shrinkage and selection operator (QWLasso) model using quaternion l_1 minimization. To solve the QWLasso model, we develop the novel quaternion fast iterative shrinkage-thresholding algorithm (QFISTA) algorithm. The fusion of the quaternion wavelets, which promote sparsity, and the sparse representation model guarantees the convergence of the proposed method to high accuracy solution.

In the third stage of the study, we combine SRWC and neural network (NN) to overcome the existing drawbacks of both approaches. More precisely, an effective convolutional autoencoder (CAE) model is proposed with the help of sparse representation in the wavelet domain in order to classify labeled images. For that, we call this method CAE-SRWC. This work is completed under a collaboration with a Master student. In the proposed approach, the CAE involves a sparse latent layer that learns the sparse codes of wavelet features. Then, a residual-based probabilistic criterion is used to assign labels to test samples based on the estimated sparse codes. Moreover, the proposed method explicitly shows a substantial reduction in the number of network parameters comparing to recent NNs.

The efficiency of the above theoretical advancements and novelties are experimentally validated by applying them on commonly used datasets, such as face and object, and comparing their results with state-of-the-art methods in the field including NNs.

In the last stage of the work, we demonstrate the capabilities of the proposed algorithms including SRWC, SRCQW, and CAE-SRWC for medical image processing with the application to skin lesion image classification. The obtained results show the potential of the newly developed methods, to classify dermoscopic skin lesion images. Moreover, the three proposed approaches show their superiority in recognizing melanoma images with high sensitivity results.

Keywords: Classification; Sparse Representation; Discrete Wavelet Transform; Quaternion Algebra; Quaternion Wavelet Transform; Skin Lesions.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Emmanuel Viennet, and co-supervisor Assoc. Prof. Marie Luong, for their help, support, and patience through my Ph.D. I thank them for choosing such an interesting thesis topic for me at the beginning of my thesis study and giving me enough independence in pursuing my ideas.

I would also like to thank my collaborators, Prof. Nikolay Sirakov and Prof. Thuong Le-Tien, for their great support and numerous suggestions throughout my Ph.D. Thanks to Prof. Nikolay Sirakov for suggesting me using the Algebra of Quaternions in the project. I am grateful to them also for giving useful comments and questions during my defense of the thesis.

I would like to thank Prof. Patrick Siarry (University of Paris-Est Créteil), Prof. Mouloud Adel (Aix Marseille Université (AMU)), for taking the time to read this thesis, giving useful comments and questions and approving the thesis to be presented.

Thanks are extended to Prof. Su Ruan (Université de Rouen) as a examiner for giving useful suggestions and questions during my defense of the thesis.

I appreciate all members of the L2TI laboratory of the University Sorbonne Paris Nord for their friendliness and support for my Ph.D.

I thank all my Vietnamese friends in Saigon (Quan Dung, Thien Tu) and in Paris (Ngoc Phuong, Hoan-Huyen, Viet-Y, Thai-Hoa, Thi-Trang, Hieu-Tran, Trong-Chau, Manh Truong, Hoang Gia, Ngoc Tuan, Bao Duy, Phuoc Nhat, Phuong Thuy, Nhat Thien, Truong Son, Thanh Yen, Trieu Theu, and my roommate Gia Khanh) for the enjoyable moments that we have had.

I would like to thank my big family in Vietnam for their support during my study.

To my dad, mom, and brother, this thesis is dedicated to you.

Contents

Abbreviations

1D	1 Dimensional
2D	2 Dimensional
3D	3 Dimensional
AE	Aautoencoder
AQ	Agebra of Quaternions
ADMM	Alternating Direction Method Multipliers
BoF	Bag-of-features
CAE	Convolutional Aautoencoder
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CS	Compressed Sensing
DCT	Discrete Cosine Transform
DOS	Data space Over Sampling
DWT	Discrete Wavelet Transform
FFT2	2D Fast Fourier Transform
HT	Hilbert Transform
kNN	k-Nearest Neighbors

LBP	**L**ocal **B**inary **P**attern
NN	**N**eural **N**etwork
PCA	**P**rincipal **C**omponent **A**nalysis
QW	**Q**uaternion **W**avelet
QWT	**Q**uaternion **W**avelet **T**ransform
ROS	**R**andom **O**ver **S**ampling
SGD	**S**tochastic **G**radient **D**escent
SR	**S**parse **R**epresentation
SRC	**S**parse **R**epresentation-based **C**lassification
SRCQW	**S**parse **R**epresentation-based **C**lassification in the **Q**uaternion **W**avelet domain
SRWC	**S**parse **R**epresentation **W**avelet-based **C**lassification
SVM	**S**upport **V**ector **M**achine

Notations

$\mathbb{R}$	the set of all real numbers.
$x \in \mathbb{R}$	a scalar.
$\mathbf{x} \in \mathbb{R}^m$	a vector.
$\mathbf{X} \in \mathbb{R}^{m \times n}$	a matrix.
$\mathbf{x}^T, \mathbf{X}^T$	the transpose of a vector $\mathbf{x}$ or a matrix $\mathbf{X}$.
x_i	the i entry of a vector $\mathbf{x}$.
$\mathbf{x}_i$	the i element of a vector set $\mathbf{X}$.
$\|\mathbf{x}\|_p$	the l_p-norm of a vector $\mathbf{x}$, which is defined as $\|\mathbf{x}\|_p = \left(\sum_i \|x_i\|^p\right)^{1/p}$ for $p \geqslant 1$.
$\|\mathbf{x}\|_0$	the l_0-norm, which counts the number of nonzero entries in a vector $\mathbf{x}$.
$\|\mathbf{X}\|_F$	the Frobenius norm of a matrix $\mathbf{X}$, which is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j \|x_{ij}\|^2}$.
$\|\mathbf{X}\|_{1,2}$	the $l_{1,2}$-norm of a matrix $\mathbf{X}$, which is defined as $\|\mathbf{X}\|_{1,2} := \sum_i \|\mathbf{X}(i,:)\|_2$, where $\mathbf{X}(i,:)$ denotes its *i-th* row.
$\mathbb{H}$	the algebra of quaternions.
$\dot{x} \in \mathbb{R}$	a quaternion.
$\dot{\mathbf{x}} \in \mathbb{R}^m$	a quaternion vector.
$\dot{\mathbf{X}} \in \mathbb{R}^{m \times n}$	a quaternion matrix.

CHAPTER 1

Introduction

In this chapter, we present the motivation and the objectives of the dissertation work on the problem of image classification including face/object recognition and medical objects classification (skin lesions). In particular, we will address the advantages of sparse representation and transform domain for image classification. Finally, we will present our main contributions with classification methods based on Sparse Representation in some transform domains.

Chapter 1 is structured as follows.

Chapter content

1.1 Motivation of the study

Image classification is an important problem in pattern recognition, computer vision and machine learning. Its role is essential for an intelligent system to exploit and manage efficiently the visual information. In the last two decades, the developments of

modern technologies have led to effective techniques both in the theory and practice of image classification in a variety of domains including healthcare, security, entertainment, financial services, and manufacturing (Aggarwa 2014; Bishop 2006; Duda et al. 2001; Zhu et al. 2020). For instance, in some medical applications, image analysis is primordial for diagnosis decision, and images represent the most important data used in the medical field. Because doctors are prone to stress, or fatigue, intelligent systems developed with the help of machine learning techniques can classify and analyse visual data automatically and more quickly than humans, hence assisting physicians in their tasks. Particularly, classification has a crucial role in computer-aided diagnosis (CAD) system. Such systems allow to detect and identify abnormalities, helping doctors make accurate diagnoses and appropriate treatment (Anthimopoulos et al. 2016; H.-D. Cheng et al. 2003; Gonzalez-Diaz 2018; Nalband et al. 2016; Suzuki 2013; Verma et al. 2016; Yanase et al. 2019; Zhou et al. 2015). Image classification has been proving capable of providing valuable cancer-fighting benefits, by classifying for example breast lesions (Zhou et al. 2015) or skin lesions (N. Codella et al. 2015; Gonzalez-Diaz 2018; Mishra et al. 2016) as either benign or malignant. Also, CAD systems play a vital role for early detection of diseases that can be beneficial for achieving better patient outcomes. In the case of skin lesion melanoma which is of our concern in this thesis, the mortality rate significantly drops if melanoma is detected and treated early. Melanoma of the skin is among the most commonly occurring cancer in the world with nearly 300,000 new cases in 2018. In particular, France is ranked 15th in the top 20 countries with the highest rates of melanoma of the skin in 2018 (Bray et al. 2018). According to (Defossez et al. 2019), the number of incident cases of melanoma in men has almost multiplied by 5 (+ 371%) between 1990 and 2018, while those for women has almost been tripled (+ 189%) in the same period. It is evident that new efficient melanoma diagnostic systems developed for clinical use are necessary to improve the survival rate.

Classification aims at assigning classes to objects/ samples, by making use of pattern recognition. Generally, classification is based on knowledge about objects and their classes. Knowledge should be represented in suitable form to describe objects. A good knowledge representation is the most important ingredient to the success of a classification method. When information is available, object classification is possible by extracting useful information about the object from its data. Typically, given a labeled training dataset consisting of two or more categories/classes, the problem is

to identify a new observation into the correct category/class. A general classification process includes two phases, a training phase where a learning model is constructed and a classification phase where the model is used to predict class labels for given test data.

More precisely, in the training phase, a set of features is first extracted based on a model for data description, also referred to as feature generation model, from the input training data. The construction of this model consists in choosing a suitable set of properties which describe some characteristics of the input data. In other words, these properties form the description features of the data. Hence, based on a feature generation model, the input data is described by a set of features (or feature vectors/ patterns). After the data representation phase through these extracted features, the relationship between the set of features and correct class label information is learned to build the classifier. In fact, classifier can be based on some optimality criterion such as the minimum error criterion which respects the value of the loss caused by classification. And a helpful approach to find out the optimal classifier setting is learning from a set of examples. Classifier learning actually enables to set classification parameters based on the training dataset, which is a set of samples (represented by their features) and their associated classes.

In the classification phase or testing phase, the features which represent an unlabeled test sample are extracted and entered as input for the classifier to assign a label to the target sample.

Generally, the larger the training set, the better the settings of the classifier. On the other hand, the time required for classification will increase with the size. Another challenge is related to the data description. Indeed, the quality of a classifier closely depends on the quality of available information. Good features enable to improve learning performance. The irrelevant features can result in worse classification accuracy. In that respect, the description of the samples should be as complex as possible. However, this results in a large number of description features as it is the case for images. Clearly, finding a good representation or extracting meaningful information and features from large and complex data is a challenging task.

In the image context, numerous feature extraction and classification methods have been developed (Aggarwa 2014; Bishop 2006; Duda et al. 2001). Many efficient features include HOG (Histogram of Oriented Gradients) (Dalal et al. 2005), SIFT (Scale Invariant Feature Transform) (Lowe 2004), SURF (Bay, Ess, et al. 2008; Bay, Tuytelaars,

et al. 2006), spatial pyramid matching (Lazebnik et al. 2006), Gabor filters (W. Li et al. 2014), the Haar wavelets (Ngo et al. 2018). Some dimensionality reduction techniques are popular feature extractions such as Principal Component Analysis (PCA) (I. T. Jolliffe 1986) and Linear Discriminant Analysis (LDA) (Balakrishnama et al. 1998). In these methods, features are projected into a new feature space with lower dimensionality, which helps reduce the complexity of the classification system and boost its speed. More recent techniques include representation learning techniques such as sparse representation (H. Cheng et al. 2013; Elad 2010; Olshausen and Field 1996; Z. Zhang et al. 2015) and deep neural networks (Georgiou et al. 2020; Hinton and Salakhutdinov 2006). If the data description is suitably chosen and the data are linearly separable, similar objects result in the proximity of their features in the feature space. Consequently, the corresponding classes can be separated in the feature space. Each feature vector represents only samples from one class. However, most of classification problems are not linearly separable due to the complex structure of data and the presence of high level of noise as well as occlusions. A linear classifier can not perfectly distinguish different classes and classify samples correctly. Researchers have shown increasing interest in dealing with the case of non-linearly separable data and developed more advanced and robust classification systems. Some efficient state-of-the-art classification methods include k-nearest neighbors (kNN) (L. Ma et al. 2010), SVM (Chapelle et al. 1999; Hearst et al. 1998), Decision tree (Quinlan 1986), Random Forest (Bosch et al. 2007), Neural Networks (Hornik et al. 1989; Krizhevsky et al. 2017; W. Zou, Lo, et al. 2006), and Deep Learning (T.-H. Chan et al. 2015; Hinton, Osindero, et al. 2006; Yuexiang Li et al. 2018) recently.

1.2 Objectives

A good classifier should achieve the generalization property, i.e. not only could well discriminate the training samples among classes but also could well represent the test samples. However, it is always challenging to achieve this property in real-world classification applications. The objective of this work is to investigate effective methods for image classification in order to address the following challenges:

- **Numerous practical problems face the issue of inadequate data or lack of data.** From the viewpoint of probability, the signal classification problem can be considered as a maximum likelihood estimation (MLE) problem. Signal identity

is decided with empirical estimates of the true densities that are learned from the training samples. A classification system may badly estimate the density, leading to a poor classifier without generalization property due to the lack of training data.

- **High dimensionality of data.** High dimensionality of data can make the above problem more harsh. In order to keep a classification system stable, the amount of required data needs to enlarge exponentially with the dimensionality. Unfortunately, high dimensionality is usually a problem encountered in image classification problems. For example, a small gray-scale image of size 100×100 has the dimension of 10000. In reality, various kinds of images (such as color, medical images...) have even much higher dimensionality and multiple channels. Hence, inadequate data and high dimensionality issues attract great interest in image classification.

- **Complex data corrupted with noises and occlusions.** Most of classification problems are not linearly separable due to the complex structure of data, such as images with intricate structures and specific properties, and the presence of occlusions. For example, in the case of human face images, samples from a same class may include large variations if the images were taken under different conditions such as illumination, viewpoint, occlusion. Another difficulty is due to the fact that image data are often prone to a high level of noise. Noises are mostly caused by the imperfection in the measurement acquisition systems and the source of the data itself. A linear classifier can not perfectly distinguish different classes and classify correctly such samples.

The success in the classification of any complex data relies on the ability of a representation learning method to reveal the meaningful features hidden in the images. Therefore, the objective of this dissertation work is to find an efficient data description and a classifier learning method for extracting meaningful knowledge from such complex and large data.

This challenge can be addressed by finding a suitable representation learning method that can capture the meaningful properties of the images. Such a method is the Sparse Representation (SR) which has achieved state-of-the-art performance in signal and image processing (Elad 2010). Over the last few decades, sparse representation has achieved

great success with successful applications in computer vision, machine learning, and signal/image processing. In the field of signal/image processing, sparse representation also emerges as a powerful tool for both theoretical and practical applications, ranging from image denoising, image inpainting (Aharon et al. 2006b), compressed sensing (Donoho 2006), super-resolution (Elad, Figueiredo, et al. 2010), image segmentation (Spratling 2013) and more recently, in image classification (Jiang et al. 2013; Wright, Y. Ma, et al. 2010; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010).

Interestingly, SR was recognized as being a primary mechanism used in the early stages of visual cortex (Olshausen and Field 1996) and considered as a main principle to efficiently represent complex data. Hence, SR has been shown to be one of the most efficient approaches producing compact as well as simple representation of the signal through only a small number of meaningful features (Olshausen and Field 1997). Such SR-based approach is known as one of the most efficient classification ones with the advantages of providing high robustness to noise and to other kinds of degradation (Olshausen and Field 1997). Sparse representation-based classification (SRC) method was initially proposed by (Wright, A. Y. Yang, et al. 2008) and based on above statements for robust face recognition. Thereafter, SRC was adapted to various classification problems, such as hyperspectral SRC (Chen et al. 2013), medical SRC (Srinivas et al. 2014), and others (Dao et al. 2016; M. Yang, L. Zhang, J. Yang, et al. 2010; Haichao Zhang et al. 2012).

The success of SRC-based methods comes from the theoretical fact that sparsity frameworks are robust to noise, occlusion, and corruption by exploiting the fact that these errors are often sparse in the standard basis (Wright, A. Y. Yang, et al. 2008). In addition, the insufficient training problem can be dealt by exploiting prior knowledge of signals as regularization terms or sparsity constraints, which capture signal relationships, in the optimization process (Srinivas et al. 2014; Haichao Zhang et al. 2012).

While many SRC methods perform in the spatial domain, we are motivated by exploiting the SR of the features in transform domains, such as wavelet domain, for enhancing the sparsity level of features and learning a compact representation of the data. Hence, we propose in this dissertation, novel sparsity frameworks in the wavelet and quaternion wavelet domains for numerous classification applications, including skin lesion classification, face identification, and object classification. In particular, we are interested in one of the most common of all cancers, which is cancer of the skin. In

particular, melanoma is one of the most dangerous skin cancers causing a majority of skin cancer deaths. According to the American Cancer Society[1] [2], the estimated number of new cases of melanoma skin lesion in the United States for 2021 is about 160,110 and correspond to the $5th$ position behind those of the colorectum, lung and bronchus, prostate, and breasts. The proposed methods will be applied to the skin lesion classification and compared with several existing methods including neural-networks based methods.

The next section of this chapter will present contributions and organization of this dissertation.

1.3 Contributions and dissertation organization

1.3.1 Main contributions

This thesis introduces three novel image classification based on SR in the transform domains with application to classification of melanoma skin lesion. The advantage of using the above novel methods is that they provide increased classification statistics, as one may observe in sections 3.4, 4.6, 5.4, and 6.3. The three methods are:

i. Sparse Representation Wavelet based Classification (SRWC)

ii. Sparse Representation Classification in the Quaternion Wavelet Domain (SRCQW)

iii. Convolutional Autoencoder Sparse Representation Wavelet Classification (CAE-SRWC)

1.3.2 Dissertation organization

An overview of the main outline and the main contributions of this dissertation is presented below. Publications related to the contribution of each chapter are also listed, as the case may be.

Chapter 2 presents the image classification state-of-the-arts, including sparse representation and deep learning based methods.

In **Chapter 3**, we propose to improve the conventional SRC method by exploiting sparsity coding in the wavelet transform domain. For this reason, the proposed method is called Sparse Representation Wavelet based Classification (SRWC). The proposed

[1] https://cancerstatisticscenter.cancer.org/?_ga=2.157868154.1341858780.1620928601-1416783839.1620928601#!/

[2] https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

method takes the advantages of wavelets which promote sparsity and provide structural information about the image, to enhance the classification performance. In this method, we fuse the image features described by the complementary information from the low sub-band of the wavelet coefficients and sparse representation to outperform the conventional SRC in terms of accuracy. To validate the capabilities and underline the advantages of the proposed SRWC, we conducted an extensive number of experiments using publicly available datasets including Extended Yale B (Georghiades et al. 2001), AR face (Martinez 1998), and COIL-100 (Nene et al. 1996) and compared our results with contemporary methods.

The material in this chapter was presented at the 2018 IEEE International Conference on Image Processing.

Chapter 4 develops another novel method for multi-class image classification based on the sparse representation (SR) approach, which operates in the Quaternion Wavelet (QW) domain. In this method, we only use features described by the information from the low-frequency coefficients of QW to construct the sparse dictionary and the classifier in the 4D space of quaternions. To calculate the quaternion SR vector, we formulate the QW Least absolute shrinkage and selection operator (QWLasso) model using quaternion l_1 minimization. To solve the QWLasso minimization model and determine the quaternion SR vector, we develop the novel Quaternion Fast Iterative Shrinkage-Thresholding Algorithm (QFISTA). In particular, we develop in the novel QFISTA an upper bound for the QWLasso model and use the upper bound as an approximation that establishes the iterative scheme to find the quaternion SR vector. The fusion of the wavelets and the SR models in the QW domain makes the novel QWLasso method achieve high accuracy of classification. Our experimental validation was conducted on four public datasets, namely Extended Yale B (Georghiades et al. 2001), AR face (Martinez 1998), AR gender, and COIL-100 (Nene et al. 1996). The experimental results show that the proposed method yields substantial accuracy improvement over the contemporary methods in the field.

The novel method and its results were presented in a manuscript that will be submitted for review, in June 2021, to the IEEE Transactions on Image Processing (IF: 9.34).

Chapter 5 proposes a novel convolutional autoencoder (CAE) architecture for sparse representation-based image classification in the wavelet domain in order to boost

the classification performance. This method offers the advantages both from the Wavelet decomposition by using the image sub-bands as inputs to learn a compact representation of image data, and the sparsity representation of the generated features to efficiently capture the meaningful characteristics of this data. This work has been conducted in collaboration with Sy NGUYEN, a Master student, in the framework of his Master Internship in the L2TI Laboratory. In the proposed approach, the autoencoder involves a sparse latent layer that learns the sparse codes of wavelet features. A residual-based probabilistic criterion is then used to assign labels to test samples based on the estimated sparse codes. Extensive experiments have been conducted on various public datasets including two digits datasets (USPS (Hull 1994) and SVHN (Netzer et al. 2011)), three face datasets (AR face (Martinez 1998), YaleB (Georghiades et al. 2001) and UMDAA-01 (Heng Zhang et al. 2015)), one object dataset COIL-100 (Nene et al. 1996), and AR gender dataset (Martinez 1998). The obtained results revealed that the proposed method yields significant classification accuracy improvement over several recent neural networks while explicitly showing a substantial reduction in the number of network parameters.

The material in this chapter was presented at the IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP).

In **Chapter 6**, we present the applications of the proposed methods to skin lesion image classification.

In particular, the SRCQW is applied to high and mixed frequency sub-bands to evaluate the performance of the method with respect to each sub-band. The theoretical and experimental results of the novel method using high and mixed frequencies were presented in a new paper, which was submitted for a review by the Signal, Image and Video Processing (IF: 1.794) in April 2021.

In **Chapter 7**, the main contributions of this dissertation are summarized with the possible future works.

1.4 Publications

Based on the research work presented in this thesis, some papers have been published in international conferences and journals:

i. **Conferences:**

1. **Ngo, L. H.**, Luong, M., Sirakov, N. M., Le-Tien, T., Guérif, S., & Viennet, E. (2018, October). Sparse representation wavelet based classification. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, October 7-10, 2018, Athens, Greece, pp. 2974-2978.
2. T. -S. Nguyen, **L. H. Ngo**, M. Luong, M. Kaaniche and A. Beghdadi, "Convolution Autoencoder-Based Sparse Representation Wavelet for Image Classification," 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), September 21-24, 2020, Tampere, Finland, pp. 1-6, **Best Student Paper Runner-up**.

ii. **Journal papers under submission:**

1. **Ngo, L. H.**, Luong, M., Sirakov, N. M., Le-Tien, T., & Viennet, E. (2021). Skin Lesion Image Classification Using Sparse Representation in Quaternion Wavelet Domain. *Signal, Image and Video Processing Journal (Under Review)*.

CHAPTER 2

State of the art of Sparse Representation-based Classification and Deep Neural Networks

Chapter content

2.1 Introduction

This chapter reviews some of the relevant state-of-the-art references from the literature in the subject of automated image classification. We first present some of the major advances in the problem of image classification, with a special emphasis on Sparse Representation-based approaches as well as Quaternion-based Sparse Representation approaches in Section 2.2 and deep neural networks in Section 2.3 that have recently

demonstrated their outstanding compared to other architectures. While this chapter does not contribute to the core of this dissertation, it prepares the necessary background for the subsequent chapters.

2.2 Sparse Representation (SR) based Classification

SR could be mathematically described as a problem whose objective is to find the sparsest solution to an underdetermined linear system. Based on this assumption, a given image can be sparsely represented over well-chosen redundant basis vectors (called '*atoms*') from a dictionary.

In the conventional SRC method (Wright, A. Y. Yang, et al. 2008), a test sample is represented by a linear combination of a few atoms taken from an overcomplete dictionary formed by the training samples. The sparsest representation of the linear combination model is first computed via a sparsity-constrained optimisation problem over the dictionary. Then, the reconstruction residual of each class is calculated and the test sample is assigned to the class with the minimum residual. However, the weakness of the method is related to the case of complex and large datasets due to the fact that the dictionary is formed by all training samples of each class. To overcome this drawback, many methods based on compact dictionary learning have been developed (Jiang et al. 2013; Vu et al. 2017; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010). One approach is to learn a discriminative dictionary of small size from a selective dataset instead of the entire dataset, such as the Discriminative K-SVD (D-KSVD) (Q. Zhang et al. 2010) and the Label Consistent K-SVD (LC-KSVD) (Jiang et al. 2013), both being based on the theory of K-SVD model (Aharon et al. 2006b).

Moreover, some recent SR-based methods are proposed in 3D space (Yi Xu et al. 2015; C. Zou et al. 2016). In these papers, the three channels of a color image are modelled as a quaternion signal and the SR models are mapped onto a 3D subspace of the AQ. (Yi Xu et al. 2015) proposed a Quaternion Sparse Representation (QSR) model, with l_0 minimization, for color image restoration and developed a Quaternion Orthogonal Matching Pursuit (QOMP) algorithm to determine the sparse coefficients. Unlike (Yi Xu et al. 2015), (C. Zou et al. 2016) proposed an l_1-norm QSR model, resulting in the quaternion Lasso (QLasso) model, which computes the sparse vector using the Alternating Direction Method of Multipliers (ADMM) algorithm (Boyd et al. 2011). The resulting model is applied to color face recognition.

2.2.1 Problem Statement of SR

In sparse representation problem, we consider an input signal $\mathbf{y} \in \mathbb{R}^m$ and optimize the empirical cost function as follows:

$$f_n(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbf{y}^i, \mathbf{D}\right), \tag{2.1}$$

where $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_n] \in \mathbb{R}^{m \times n}$ is the dictionary whose columns are basis vectors, and ℓ is a loss function such that $\ell(\mathbf{y}, \mathbf{D})$ should be small if $\mathbf{D}$ is "good" at representing the input signal $\mathbf{y}$ in a sparse representation. Note that, in this setting, over-complete dictionaries with $n > m$ are allowed. In machine learning and image processing, sparse regularized problems consist in fitting some model parameters $\mathbf{x} \in \mathbb{R}^n$ to the training samples, while having the *a-priori* assumption that $\mathbf{x}$ should be sparse. This can be achieved by minimizing $F(\mathbf{x})$ including a smooth convex function $f(\mathbf{x})$, which is typically a data fitting or data reconstruction term in image processing, and a sparse regularization term $g(\mathbf{x})$:

$$\ell(\mathbf{y}, \mathbf{D}) \triangleq \min_{\mathbf{x} \in \mathbb{R}^n} \{F(\mathbf{x}) := f(\mathbf{x}) + \lambda g(\mathbf{x})\}, \tag{2.2}$$

where $\mathbf{x}$ is a sparse vector, and λ is a non-negative regularization parameter, which controls the trade-off between reconstruction error and regularization. To find the sparsest solution $\mathbf{x}$, l_0 pseudo-norm should be a natural choice for $g(\mathbf{x})$ with the purpose of counting the number of non-zero entries in $\mathbf{x}$. However, it is NP-hard to find the sparsest solution of Eq. 2.2 in this setting. Fortunately, greedy algorithms or convex relaxation can provide the approximate solutions. If the solution $\mathbf{x}$ is sparse enough, the solution of the l_0-minimization problem is equivalent to the solution of the l_1-minimization problem, also known as the Lasso (Tibshirani 1996):

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left[\frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1\right], \tag{2.3}$$

where $\mathbf{y} \in \mathbb{R}^m$ is an input signal and $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_n] \in \mathbb{R}^{m \times n}$ is a dictionary whose columns are the dictionary atoms. If the value of λ is large enough, $\mathbf{x}$ is known to be sparse. Thus, only a few dictionary atoms are involved in the representation. To prevent $\mathbf{D}$ from having arbitrarily large values (which would lead to arbitrarily small values of $\mathbf{x}$), it is common to constrain its columns $[\mathbf{d}_1, ..., \mathbf{d}_n]$ to have an l_2-norm less

than or equal to one, i.e. $\mathbf{D} \in \mathbb{R}^{m \times n} \quad s.t. \quad \forall j (j = 1, ..., n),\ \|d_j\|_2^2 \leqslant 1$. The problem of efficiently solving Eq. 2.3 has received a lot of attention lately. Eq. 2.3 can be rewritten as a matrix factorization problem with respect to the dictionary $\mathbf{D}$ and the sparse codes $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_k] \in \mathbb{R}^{n \times k}$, which is convex with respect to $\mathbf{X}$ while $\mathbf{D}$ is fixed:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times k}} \left[\frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2 + \lambda \|\mathbf{X}\|_1 \right], \tag{2.4}$$

where, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_k] \in \mathbb{R}^{m \times k}$ is the matrix of input signals, and $\|\mathbf{X}\|_1$ denotes the l_1 norm of sparse matrix $\mathbf{X}$ which is the sum of the its coefficients.

2.2.2 Sparse Representation-based Classification (SRC)

Recently, sparse representation-based classification (SRC) framework (Wright, A. Y. Yang, et al. 2008) has attracted the attention of the computer vision community for image classification as a remarkable contribution to its development. SRC utilize the discriminative capability of sparse representation to deal with some of the aforementioned challenges. Given a sufficient set of training samples of k categories/ classes, any new sample with a specific category can be considered as a linear combination of the training samples with the same category. By using a dictionary consisting of training samples from all categories, any new and unlabeled test sample can be sparsely represented with respect to such dictionary. It is noted that the class-specific design of the dictionary keep the sparsity assumption existing in a linear representation model. The main concepts of SRC (Wright, A. Y. Yang, et al. 2008) are presented as follows.

Consider a classification problem with k classes. Let n_c be the number of training samples from class c, for $1 \leqslant c \leqslant k$. Denote by $\mathbf{D}_c$ the set of real labeled training samples from class c, in which each columns of $\mathbf{D}_c$ is the vectorized vector of an image. The training samples in $\mathbf{D}_c$ are arranged as columns of a matrix $\mathbf{D}_c = [\mathbf{d}_{c,1}, \mathbf{d}_{c,2}, ..., \mathbf{d}_{c,n_c}] \in \mathbb{R}^{m \times n_c}$. A new test sample $\mathbf{y} \in \mathbb{R}^m$ from class c can be approximately expressed as $\mathbf{y} = \mathbf{D}_c \mathbf{x}_c$, where $\mathbf{x}_c$ is a sparse vector associated with the c-th class. Let $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_k] \in \mathbb{R}^{m \times n}$ be the matrix of all training samples of k classes with a total number of n atoms, where $n = \sum_{c=1}^{k} n_c$. Then, for any test sample $\mathbf{y}$, the objective is to correctly predict its label. The procedures of SRC are as follows:

i. Given an input sample $\mathbf{y}$ and dictionary $\mathbf{D}$, the sparse coefficients are computed via solving the l_1-norm minimization, also known as Lasso problem (Tibshirani

1996):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2.5}$$

where λ denotes the regularization parameter.

ii. Perform classification by first computing the reconstruction residual of SRC as follows:

$$r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}\delta_c(\widehat{\mathbf{x}})\|_2, \tag{2.6}$$

where $\delta_c(\widehat{\mathbf{x}})$ is the part of $\widehat{\mathbf{x}}$ whose nonzero entries are associated with the c-th class. Then, the identity of $\mathbf{y}$ is determined by the class with minimal residual:

$$c = identity(\mathbf{y}) = \arg\min_c r_c(\mathbf{y}). \tag{2.7}$$

Although SRC shows interesting results, its performance is still limited in case of high level noise and artifacts in the original training images. In addition, if a large amount of data is involved in the scheme, the computational complexity will increase because the dictionary is formed by all training samples of each class. Moreover, the choice of dictionary is important for the success of a SR based method. However, by using directly original training images as dictionary, the discriminative information that is hidden in the training images cannot be fully exploited. Therefore, dictionary learning is proposed as a solution to the aforementioned problems, at least to some extent. The next subsection will present the Label Consistent K-SVD (LC-KSVD) (Jiang et al. 2013), which is a well-known dictionary learning method to sparse representation based image classification, with impressive results in image classification.

2.2.3 Label Consistent K-SVD (LC-KSVD)

LC-KSVD learns a discriminative over-complete dictionary and an optimal linear classifier simultaneously (Jiang et al. 2013). It yields dictionaries so that feature points with the same class labels have similar sparse codes (Jiang et al. 2013). During the dictionary learning process, the label information is exploited with each dictionary item (column of the dictionary matrix) to enforce the discriminative sparse codes $\mathbf{x}$.

Next, we will present the main concept of the two methods, LC-KSVD1 and LC-KSVD2 in (Jiang et al. 2013). The latter is differentiated from the former by incorporating the classification error term in the objective function, which makes

the learning process more optimal for image classification. Denote by $\mathbf{Y}$ a set of n-dimensional N input signals, i.e. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N] \in \mathbb{R}^{n \times N}$. An objective function for learning a discriminative dictionary with K items for sparse representation of $\mathbf{Y}$ can be defined as it is done by:

- LC-KSVD1 (Jiang et al. 2013):

$$< \mathbf{D}, \mathbf{A}, \mathbf{X} >= \underset{\mathbf{D},\mathbf{A},\mathbf{X}}{\arg\min} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2 \quad s.t. \forall i, \|\mathbf{x}_i\|_0 \leqslant T, \quad (2.8)$$

  where T is a sparsity constraint factor, $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K] \in \mathbb{R}^{n \times K}$ ($K > n$, making the dictionary over-complete) is the learned dictionary, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{K \times N}$ are the sparse codes of input signals $\mathbf{Y}$, α controls the relative correlation between the reconstruction and label consistent regularization, $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_N] \in \mathbb{R}^{K \times N}$ are the 'discriminative' sparse codes of input signals $\mathbf{Y}$ for classification. $\mathbf{q}_i = [q_{i,1}, q_{i,2}, ..., q_{i,K}]^T \in \mathbb{R}^K$ is a 'discriminative' sparse code corresponding to an input signal $\mathbf{y}_i$, where $q_{i,k}$ ($k = 1...K$) equals 1 if the input signal $\mathbf{y}_i$ and dictionary item $\mathbf{d}_k$ share the same label, and 0 otherwise. $\mathbf{A} \in \mathbb{R}^{K \times K}$ is a linear transformation matrix. Here we identify a linear transformation, $g(\mathbf{x}, \mathbf{A}) = \mathbf{A}\mathbf{x}$, which transforms the original sparse codes $\mathbf{x}$ to be most discriminative in sparse feature space $\mathbb{R}^K$.

- LC-KSVD2 (Jiang et al. 2013):

$$< \mathbf{D}, \mathbf{W}, \mathbf{A}, \mathbf{X} >= \underset{\mathbf{D},\mathbf{W},\mathbf{A},\mathbf{X}}{\arg\min} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2 + \beta \|\mathbf{H} - \mathbf{W}\mathbf{X}\|_F^2 \quad s.t. \forall i, \|\mathbf{x}_i\|_0 \leqslant T, \quad (2.9)$$

  where the term $\|\mathbf{H} - \mathbf{W}\mathbf{X}\|_F^2$ represents the classification error, $\mathbf{W}$ denotes the classifier parameters, $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N] \in \mathbb{R}^{m \times N}$ are the class labels of input signals $\mathbf{Y}$, $\mathbf{h}_i = [0, 0, ..., 1, ..., 0]^T \in \mathbb{R}^m$ is a label vector corresponding to an input signal $\mathbf{y}_i$, where the non-zero position indicates the class of $\mathbf{y}_i$, while α and β are the scalars controlling the relative contribution of the corresponding terms.

The optimal solutions of $\mathbf{D}, \mathbf{A}$, and $\mathbf{W}$ are obtained by employing the efficient K-SVD algorithm (Aharon et al. 2006b) to solve Eq. 2.8 or Eq. 2.9.

Classification scheme: For a test image $\mathbf{y}_i$, given the learned dictionary $\mathbf{D}$, its

sparse representation $\mathbf{x}$ is computed by solving the optimization problem:

$$\mathbf{x}_i = \arg\min_{\mathbf{x}_i} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 \quad s.t. \|\mathbf{x}_i\|_0 \leqslant T. \tag{2.10}$$

Then the label c of the test sample $\mathbf{y}_i$ is estimated by simply using the linear predictive classifier $\mathbf{W}$ as follows:

$$c = identity(\mathbf{y}_i) = \arg\max_c (l = \mathbf{W}\mathbf{x}_i), \tag{2.11}$$

where $l \in \mathbb{R}^m$ is the class label vector.

In the next two sub-sections, we will review the two methods, namely Quaternion-based Sparse Representation (QSR) (Yi Xu et al. 2015) and Quaternion Sparse Representation-based Classification (QSRC) (C. Zou et al. 2016), that work with color images using the sparse representation and algebra of quaternions.

2.2.4 Quaternion-Based Sparse Representation (QSR)

Xu et al (Yi Xu et al. 2015) fuses sparse representation with algebra of the quaternions to represent each color image as a quaternion matrix. A quaternion-based over-complete dictionary is learned through the K-quaternion singular value decomposition (QSVD) method (Yi Xu et al. 2015). K-QSVD consistently transforms the color images to an orthogonal color space to select the sparse basis atoms during the dictionary learning process. Then, QOMP (quaternion orthogonal matching pursuit) method is exploited to compute the sparse coefficients (Yi Xu et al. 2015). In such color space, what make the proposed method outstanding are full preservation of the intrinsic color structures in the images during sparse reconstruction and the lower redundancy between the dictionary atoms of different color channels.

Concretely, a RGB color image patch is vectorized using the pure quaternion form as: $\dot{\mathbf{y}} = 0 + \mathbf{y}_r i + \mathbf{y}_g j + \mathbf{y}_b k \in \mathbb{H}^m$, where the subscript r, g, and b denote the RGB channels respectively. Accordingly, the learned quaternion dictionary and the corresponding quaternion sparse vector are represented as: $\dot{\mathbf{D}} = \mathbf{D}_s + \mathbf{D}_r i + \mathbf{D}_g j + \mathbf{D}_b k \in \mathbb{H}^{m \times n}$ and $\dot{\mathbf{x}} = \mathbf{x}_0 + \mathbf{x}_1 i + \mathbf{x}_2 j + \mathbf{x}_3 k \in \mathbb{H}^n$, where s denotes the scalar part (of a quaternion). The quaternion-based sparse representation (QSR) model is formulated as follows (Yi Xu et al. 2015):

$$\arg\min \|\dot{\mathbf{x}}\|_0, \quad s.t.\ \dot{\mathbf{y}} = \dot{\mathbf{D}}\dot{\mathbf{x}}, \tag{2.12}$$

where $\dot{\mathbf{D}} \in \mathbb{H}^{m \times n}$ is a quaternion dictionary including n pure quaternion atoms, $\dot{\mathbf{x}} \in \mathbb{H}^n$ is a quaternion sparse vector corresponding to the input signal $\dot{\mathbf{y}} \in \mathbb{H}^m$, with its components $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3 \in \mathbb{R}^n$. The l_0 norm $\|\dot{\mathbf{x}}\|_0$ counts the number of nonzero elements in $\dot{\mathbf{x}}$.

The quaternion dictionary learning process can be formulated as an extension of the QSR model in Eq. 2.12 with unknown dictionary and sparse quaternion codes as follows:

$$< \widehat{\dot{\mathbf{D}}}, \widehat{\dot{\mathbf{X}}} >= \arg\min_{\dot{\mathbf{D}}, \dot{\mathbf{X}}} \left\| \dot{\mathbf{Y}} - \dot{\mathbf{D}}\dot{\mathbf{X}} \right\|_F^2 + \lambda \left\| \dot{\mathbf{X}} \right\|_0, \tag{2.13}$$

where $\dot{\mathbf{Y}} \in \mathbb{H}^{m \times N}$ is the collection of the sample image patches, $\dot{\mathbf{X}} \in \mathbb{H}^{n \times N}$ is the sparse quaternion coefficient matrix. K-QSVD, which is an extension of the well-known K-SVD (Mairal, Elad, et al. 2007) algorithm, is developed to optimize the dictionary learning process. It consists of two main steps: sparse coding and dictionary updating.

During the sparse coding stage, QOMP algorithm is developed to find the solution to the sparse coefficient matrix $\dot{\mathbf{X}}$, given a fixed quaternion dictionary $\dot{\mathbf{D}}$ in 2.13. QOMP is a counterpart of the OMP (Pati et al. 1993) algorithm, but works with quaternion numbers. It solve the sparse representation problem of a signal $\dot{\mathbf{y}} \in \mathbb{H}^m$ on a quaternion dictionary $\dot{\mathbf{D}} \in \mathbb{H}^{m \times n}$ such that:

$$\dot{\mathbf{x}} = \arg\min_{\dot{\mathbf{x}}} \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\dot{\mathbf{x}} \right\|_2^2 \quad s.t. \|\dot{\mathbf{x}}\|_0 \leqslant T, \tag{2.14}$$

where $\dot{\mathbf{x}} \in \mathbb{H}^n$ is the sparse vector of coefficients and $\|\dot{\mathbf{x}}\|_0 \leqslant T$ is the stopping criterion. It eases the NP-hard l_0-minimization problem by specifying the maximum number of non-zero components per signal.

During the dictionary updating step, the quaternion dictionary $\dot{\mathbf{D}}$ can be learned given the sparse quaternion codes. K-QSVD shows its high efficiency by updating both the dictionary atoms and the sparse quaternion coefficients jointly. Both of the atom $\dot{\mathbf{d}}_k$ and its corresponding coefficients in $\dot{\mathbf{X}}^l$, the $l-th$ row of $\dot{\mathbf{X}}$, are updated simultaneously by decomposing the representation error $\dot{\mathbf{E}}_l = \dot{\mathbf{Y}} - \sum_{p \neq l} \dot{\mathbf{d}}_p \dot{\mathbf{X}}^p$ using QSVD.

2.2.5 Quaternion Sparse Representation-based Classification (QSRC)

Inspired by the SRC method, (C. Zou et al. 2016) proposed a novel SRC method in the quaternion space for color images classification, namely quaternion SRC (QSRC). Similar to SRC, QSRC exploit the l_1-norm minimization problem in the quaternion

space. QSRC naturally preserves the inherent correlation among the color channels of color images with quaternions. More precisely, the authors extend the SRC to the quaternion setting. Consider a classification problem with k classes. The main stages of QSRC are as follows:

i. Given $\dot{\mathbf{y}}$ and $\dot{\mathbf{D}}$, the sparse quaternion coefficients vector is computed via solving the quaternion l_1-norm minimization, also known as quaternion Lasso (QLasso) problem such that:

$$\hat{\dot{\mathbf{x}}} = \underset{\dot{\mathbf{x}} \in \mathbb{H}^n}{\arg\min} \left\{ \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\dot{\mathbf{x}} \right\|_2^2 + \lambda \|\dot{\mathbf{x}}\|_1 \right\}, \tag{2.15}$$

where λ denotes the regularization parameter. Eq 2.15 can be solved applying the Alternating Direction Method of Multipliers (ADMM) framework (Boyd et al. 2011).

ii. Perform classification by first computing the residual of QSRC as follows:

$$r_c(\dot{\mathbf{y}}) = \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\delta_c(\hat{\dot{\mathbf{x}}}) \right\|_2, \tag{2.16}$$

where $\delta_c(\hat{\dot{\mathbf{x}}})$ is the part of $\hat{\dot{\mathbf{x}}}$ whose nonzero entries are associated with the c-th class. Then, the identity of $\dot{\mathbf{y}}$ is determined to the class with minimal residual:

$$c = identity(\dot{\mathbf{y}}) = \arg \min_{c=1,\ldots,k} r_c(\dot{\mathbf{y}}). \tag{2.17}$$

From the application perspective, the sparsity principle has had great effect in several domains, especially in image processing (Elad and Aharon 2006; Mairal, Bach, et al. 2014).

2.3 Deep Neural Networks

In this section, we review some of the key contemporary studies in deep neural network (NN) domain, especially when applying to image classification. This review prepares the necessary background for the study in chapter 5, where deep neural networks can be fused with sparse representation to enhance the capability of classification systems. In recent years, deep neural networks, e.g. convolutional or recurrent neural networks, have become one of the most popular and trendy approaches in computer vision (Krizhevsky

et al. 2017; Simonyan and Zisserman 2014; Szegedy, Liu, et al. 2015; Taigman et al. 2014). In fact, this concept was proposed a few decades ago (Y. LeCun et al. 1998), while benefiting from the huge amounts of available data as well as the great power of computers (e.g. powerful GPUs) nowadays. They allow deep networks for learning a large number of model parameters and better representing the images (Jaderberg et al. 2015; Krizhevsky et al. 2017; Simonyan and Zisserman 2014). In the context of image classification, convolutional neural networks (CNNs) (Y. LeCun et al. 1998) might be the most popular models recently. With the help of the linear operations constrained to be local convolutions and a down-sampling operation in the feature pooling layers, CNNs are capable of modeling the local stationarity in images and combining the low-level and high-level features (Zeiler et al. 2014). CNNs has started a revolution in the practice of computer vision. However, they are prone to over-fitting and data-hungry. They also suffer from vanishing and exploding gradients (Georgiou et al. 2020). In the other words, these networks still suffer from many limitations. A number of research with different manners have been proposed to resolve these issues. More precisely, these researches studies various elements of CNNs such as the activation functions, the normalization layers, training strategies, or the network architecture (Georgiou et al. 2020) (e.g. the inception networks (Szegedy, Ioffe, et al. 2017))

Deep models build their deep visual representations from the available data with no specific priors required. Thanks to available big data and powerful GPUs, deep networks have achieved outstanding performance in natural image classification (ImageNet) (Krizhevsky et al. 2017; Simonyan and Zisserman 2014; Szegedy, Liu, et al. 2015), face recognition (Parkhi et al. 2015; Taigman et al. 2014), and fine-grained classification (Jaderberg et al. 2015; Wah et al. 2011). We briefly present three of the main operations used to build the data representation in deep models as follows:

- Convolution: Given a three-dimensional feature map $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$, the convolution layer calculates outputs $\mathbf{y} \in \mathbb{R}^{H'' \times W'' \times D''}$ as the convolution between $\mathbf{x}$ with D'' learned filters $\mathbf{f} \in \mathbb{R}^{H' \times W' \times D}$ as follows (Vedaldi et al. 2015):

$$y_{i'',j'',d''} = b_{d''} + \sum_{i'=1}^{H'} \sum_{j'=1}^{W'} \sum_{d=1}^{D} f_{i',j',d} x_{i''+i'-1,j''+j'-1,d}, \tag{2.18}$$

  where $b_{d''}$ denotes the bias; $H'' = 1 + H - H'$ and $W'' = 1 + W - W'$ if unpadded convolution is applied with a stride of 1.

- Rectification: Modern deep CNNs use a half-rectification activation function, also known as rectified linear unit (ReLU), defined by:

$$\mathbf{y} = \max(0, \mathbf{x}). \tag{2.19}$$

This simple non-linearity has been shown to provide significant advances when compared to traditional sigmoid activation functions (Glorot et al. 2011; Maas et al. 2013), and is also tightly related to sparse coding (Fawzi et al. 2015).

- Pooling: A pooling operation aims to provide invariant to the classifier through the computation of summary statistics over the discriminative features. Given a feature map x, the pooled representation is given by:

$$y_{i'',j'',d} = P\left(\left\{x_{i''+i'-1,j''+j'-1,d}\right\}_{\substack{1\leqslant i'\leqslant W' \\ 1\leqslant j'\leqslant H'}}\right), \tag{2.20}$$

where W' and H' are the width and heights of the pooling regions, respectively and P denotes the pooling operator (e.g. average pooling, max pooling, etc (Boureau et al. 2010)). A pooling operation often go along with a sub-sampling of the feature channel.

More common elementary operations for CNN architectures can be seen in (Vedaldi et al. 2015). Figure 2.1 shows a simple CNN architecture. Given the training images and the specific architecture, the CNN is then trained in an end-to-end manner, where the convolutional filters $\mathbf{f}$ are learned and updated after each iteration. Generally, stochastic gradient descent (SGD) optimization, which exploits backpropagation (Y. A. LeCun et al. 2012) to compute the gradients, are employed to train CNNs. In the recent years, some improving optimization algorithms have been developed (Kingma et al. 2014; Martens 2010; Martens and Grosse 2015), which outperforms the existing optimization ones.

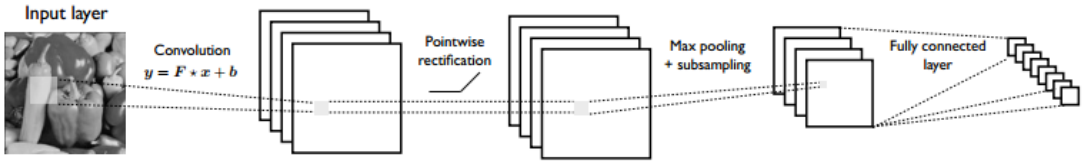


Figure 2.1: *Main structure of a CNN with stacked series of linear and nonlinear operations. (Fawzi 2016)*

Although deep CNNs have achieved impressive successes, the theoretical questions

regarding to its internal mechanisms leading to these magnificent results still needs to be answered. The authors of (Mahendran et al. 2015) have proposed visualization tools to understand the representations of features learned by deep CNNs. These visualization strategies come up with some empirical understanding about the layers where the representation invariant is achieved (Mahendran et al. 2015). While (Simonyan, Vedaldi, et al. 2013) and (Zeiler et al. 2014) have proposed different visualization strategies, which attempt to maximize the activation of single neurons. It is shown with these visualization strategies that neurons in lower layers tend to be sensitive to edges, while neurons in higher layers are sensitive to semantic objects with similar visual appearance.

Up until now, there is a lack of principled methodology and theory for deep learning, specifically CNNs. Deep models are usually considered as black boxes since they were originally proposed at the end of the 70's (Fukushima 1979), then were developed in the 90's (Y. LeCun et al. 1998) and recently are evolved deeper (Krizhevsky et al. 2017; Simonyan and Zisserman 2014). Some of important problems regarding the CNNs models can be enumerated as follows:

i. **Model regularization.** Deep models are mainly regularized based on early-stopping the optimization procedure, model averaging (Srivastava et al. 2014; Wan et al. 2013), and data augmentation (Ciresan et al. 2012). Nevertheless, model averaging approaches, such as Dropout (Srivastava et al. 2014), are deficiently understood from a theoretical perspective (Wager et al. 2014) and seem to weaken benefits (Wan et al. 2013), while data augmentation is a powerful technique that helps lessen the importance of regularization by artificially producing virtual examples to increase the amount of the training samples. However, such generation needs domain-knowledge, which is not always available, and manual selection of various hyper-parameters.

ii. **Unsupervised learning.** Deep neural networks have successfully solved the tasks with given labeled training samples. But such data is not always available. Although unsupervised deep learning approaches such as autoencoders (AEs) (Hinton, Osindero, et al. 2006; Hinton and Salakhutdinov 2006) or restricted Boltzmann machines (Smolensky 1986) have had mixed success. In particular, an AE is a symmetric NN that learns the data features in an unsupervised manner. Bourlard and Kamp (Bourlard et al. 1988) presented its very first basic version

as an auto-association network, whose objective is to decrease the data dimensionality by using a fully connected layer. Then the network tries to reconstruct the input data from the intermediate representation, which should carry most of the information of the input. Later, the AE as widely known now was proposed by Hinton and Salakhutdinov (Hinton and Salakhutdinov 2006), in which the network consists of multiple layers. Specifically, the AE consists of an encoder, a hidden layer, and a decoder. The encoder nonlinearly maps input data to a latent representation through an activation function (e.g., sigmoid or ReLU). Then the latent representation is mapped linearly onto the decoder to reconstruct an approximation of the input data through an activation function. The training process of an AE is conducted through the back-propagation algorithm to minimize the reconstruction error between the input and reconstructed data, which yields optimal parameters of the network. Since the development of AE, the sparse autoencoder was one of its first variants, whose purpose is to transform the data into a higher dimensional space through a sparse representation. Such representation could make the data be linearly separable (Georgiou et al. 2020) and allows a straightforward interpretation of the data by a small number of hidden features (Ranzato et al. 2007).

Deep models need to be improved in unsupervised learning, which is related to the question of how to regularize them.

iii. **Functional spaces and properties of deep networks.** It is essential to understand the geometry of functional spaces corresponding to deep neural networks because it might answer the issue of regularization. It provides solutions to manage the variations of prediction function in a principled fashion. Recently, (Bruna et al. 2013) endeavor to fill in this gap with the help of scattering transform where CNNs (Y. LeCun et al. 1998) are fused with the wavelets. The theory of scattering transform provides interesting insight into invariant properties of image representations constructed by deep neural networks. Differing from other deep networks, scattering transform exploits the predefined wavelet functions instead of the learned filters in the CNNs, which may limits the performance of the networks. Other studies have investigated the properties of deep networks under the assumption of independent identically distributed random weights (Giryes et al. 2016). More precisely, (Giryes et al. 2016) study the three fundamental prop-

erties of deep networks. Firstly, the metric information of input data is preserved when propagating through the layers of deep neural networks. This allows firmly recovering the data from the obtained features. Secondly, information of unseen data can be carried by the training samples through the learning process. Finally, the deep networks have the ability to differently treat in-class and out-of-class data from different classes (Giryes et al. 2016). Similar works should be studied to clearly understand the functional spaces and properties of deep networks.

iv. **Optimization.** Deep models are often formulated as the minimization problem of a non-convex objective function. Hence, it is not possible to find the global optimum in general, which make it difficult to analyze the model. Recently, some significant progressions from an optimization point of view have been made with a few accomplished theoretical results suggesting that very deep networks can be controllable under some assumptions (Choromanska et al. 2015; Livni et al. 2014). Nevertheless, there is still a gap between theory and practice, and the theory does not give an algorithm on how to successfully design deep architectures.

2.4 Conclusion

In this chapter we have presented two parts of related research on Sparse Representation based Classification (SRC) and Deep Neural Networks. In the first part, image classification with the help of Sparse Representation based approaches is introduced, including the conventional SRC method and its variations (LC-KSVD, QSR, and QSRC). While the second part review some key studies of neural networks in image classification, which is crucial for the study in chapter 5. In the next chapter, our first proposed work based on the Sparse Representation based Classification will be presented.

CHAPTER 3

Contribution to the Sparse Representation Classification in the Wavelet Transform Domain

Chapter content

3.1 Introduction

If the recent years have witnessed an explosion of interest toward machine and deep learning, which helps computers learn from data without being explicitly programmed, sparse representation (SR) modeling has also undergone strong expansion with applications in machine learning, image processing, statistics and computer vision. SR has achieved state-of-the-art frameworks in both theoretical research and practical applications in the field of image processing, such as denoising (Elad and Aharon 2006), inpainting (Aharon et al. 2006b), super-resolution (Elad, Figueiredo, et al. 2010), segmentation (Lu et al. 2014; Moradi et al. 2019; Spratling 2013; Unser 1995) and classification (Chen et al. 2013; Jiang et al. 2013; Wright, Y. Ma, et al. 2010; Wright, A. Y. Yang, et al. 2008; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010). The success of SR can be explained by the fact that it behaves like a primary mechanism used in the early stages of visual cortex (Olshausen and Field 1996). Hence, it provides a useful tool to efficiently represent complex data with a compact and simple interpretation of the signal through only a small number of important features (Olshausen and Field 1997). Moreover, sparsity-based approach is efficient for its robustness to noise, occlusion, and corruption (Wright, A. Y. Yang, et al. 2008). Mathematically, SR consists of finding the sparsest solution to an underdetermined linear system. In other words, SR locates the solution with the fewest nonzero entries. Based on the observation that small-scale structures are inclined to repeat themselves in a single image or a group of similar images (Elad 2010), an image can be sparsely represented over some well-chosen redundant basis. SR is related to compressed sensing (CS) (Candès et al. 2006; Donoho 2006). Donoho (Donoho 2006) first proposed the original notion of CS. According to CS hypothesis, if a signal is sparse enough, the original signal can be reconstructed by utilizing a few measured values. (Candès et al. 2006) proved that the original signal could be accurately reconstructed by utilizing a small amount of Fourier coefficients. Thus, a large number of algorithms based on CS hypothesis have been developed to address a number of problems in various fields including SR, encoding measuring, and reconstructing algorithm. In (Elad, Figueiredo, et al. 2010; Wright, Y. Ma, et al. 2010), it is proven that the SR theory is one of the most outstanding techniques used to solve problems in denoising, face recognition/classification, pattern recognition/classification, and computer vision.

Classification is a typical task in supervised learning. A fundamental problem in

supervised classification is to use labeled training samples from k distinct object classes to predict the category to which a new observation belongs. From the perspective of the way of exploiting "atoms," SRC can be sorted into two major categories: holistic representation based and local representation based methods (Z. Zhang et al. 2015). Holistic representation based methods exploit the training samples of all classes to represent the test sample, whereas local representation based methods only utilize training samples of some classes. Most existing SRC algorithms belong to the holistic representation based group (Jiang et al. 2013; Wright, A. Y. Yang, et al. 2008; Q. Zhang et al. 2010).

In this chapter, in order to achieve high accuracy and address the challenging problem of large database, we investigate a new approach of SRC by exploiting sparsity coding in the wavelet transform domain. For this reason, the proposed method is called Sparse Representation Wavelet based Classification (SRWC). The proposed SRWC can be considered as a holistic representation based method. The proposed method takes the advantages from: i) the wavelet decomposition which promotes sparsity and provides structural information about the image data, ii) the dimensionality reduction method using PCA for reducing the complexity of the problem, and iii) the sparse representation of the generated features to efficiently capture the useful characteristics of this data.

After the wavelet decomposition, the low-frequency image sub-band information is projected into a new feature space with lower dimensionality using PCA. Taking advantages of the generated features, we build an overcomplete dictionary, which allows for representing a test sample from a given dataset. Hence, the test samples are considered as a linear combination of the transformed training samples into the wavelet domain. This representation is naturally sparse, and help to reject test samples, which do not belong to the dataset (Wright, A. Y. Yang, et al. 2008). Then, the test sample features are sparsely coded for the classification step, which is based on the minimum reconstruction residual. To validate the capabilities and underline the advantages of the novel SRWC, we conducted an extensive number of experiments using publicly available datasets and compared our results on face and object classification with several contemporary methods. The results demonstrated that the proposed approach outperforms state-of-the-art methods.

The rest of the chapter is organized as follows: Section 3.2 introduces the related works; Section 3.3 presents the wavelet transform and develops the novel method,

SRWC; In section 3.4, SRWC is validated on commonly used datasets, and compared with several contemporary methods; Section 3.5 concludes the chapter while listing the contributions.

3.2 Related works

SRC method (Wright, A. Y. Yang, et al. 2008) assumes that a test sample can be represented as a linear combination of a few basis vectors taken from a dictionary whose base elements are the training samples. More specifically, SRC exploits the linear combination of training samples to represent the test sample by computing the sparse codes of the test sample on the dictionary basis. The reconstruction residuals of each class are computed through the SR coefficients and training samples. The membership of a test sample is determined by the minimum residual. In (Wright, A. Y. Yang, et al. 2008), it is shown that corrupted face images could be recognized by the SRC algorithm, developed for robust face detection. Later, SRC was adapted to numerous image classification problems, such as hyperspectral SRC (Chen et al. 2013).

According to recent studies (Jiang et al. 2013; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010), instead of using all the training samples as a dictionary, learning a dictionary from them could effectively improve the SRC performance. Based on the theory of K-SVD model (Aharon et al. 2006a), discriminative K-SVD (D-KSVD) (Q. Zhang et al. 2010) and label consistent K-SVD (LC-KSVD) (Jiang et al. 2013) are constructed to learn a discriminative dictionary, where the sparse codes are projected to be sparse enough. In (Jiang et al. 2013), the authors differentiated LC-KSVD2 from LC-KSVD1 by including the classification error term in the objective function for dictionary learning, which makes the dictionary optimal for the classification task. In (M. Yang, L. Zhang, Feng, et al. 2014), the FDDL algorithm employs Fisher discrimination criterion to construct dictionaries and sparse codes.

The SRC methods mentioned above are applied to the spatial domain, to construct a dictionary used in classification. To improve the image classification performance, transform domains could be considered as a promising tool. In (Ghazali et al. 2007), the authors utilized the wavelet transform as an extractor to get the high-leveled feature, which is then used to classify narrow and broad weeds. Later, in (Huang et al. 2008) the wavelet transform is used to extract the spectral and spatial features of very high resolution (VHR) satellite imagery. Then these features are fed to a support vector

machine (SVM) to classify the VHR satellite imagery. (Jian et al. 2009) proposed to use Gabor wavelet features for image texture classification. Moreover, they developed feature selection functions based on the Fisher discrimination criterion to choose the features that helps to better discriminate the images between classes. In (Tian et al. 2018), the authors fused frequency domain features, extracted by the Fast Fourier transform (FFT), and the sparse representation or collaborative representation to classify images, which is proved to be efficient and robust. (S. Zhang et al. 2012) extracted the image features using Gabor wavelets. Then these features were fed to SRC, K-nearest neighbor (KNN), support vector machines (SVM), and artificial neural network (ANN) to evaluate their efficiency in a facial expression recognition task.

Based on above observations, in our approach, we exploit the advantages of utilizing both the wavelets, which are naturally sparse and provide structural information about the image, as well as the sparse representation of the resulting features to efficiently capture the important characteristics of this data, for the classification task. Hence, the fusion of the wavelet coefficients and SR helps to enhance the classification accuracy.

3.3 Proposed method: Sparse Representation Wavelet Based Classification (SRWC)

The advantages of the proposed sparse-representation based SRWC method are that it is performed in the wavelet domain, promoting sparsity and leading to better discrimination and hence improving the classification accuracy. Indeed, the proposed Sparse Representation based approach allows not only learning a compact representation of images data by using the wavelet coefficients as features, but also capturing the meaningful characteristics of this data. Furthermore, it was proven, as described in (W. Zou and Yan Li 2007), that the extracted low-pass sub-band coefficients lead to better discrimination. Another advantage is that feature extraction of wavelet coefficients followed by the dimension reduction PCA method helps to reduce data dimension and computational cost.

Consider the classification problem of k classes. As proposed in (W. Zou and Yan Li 2007), image features can be underlined by projecting the distribution of wavelet coefficients onto the x and y-axes. These projections can be represented by histograms with eight bins in both the x and y-axes. It is also shown that features described by wavelet coefficients can significantly improve the image classification. However, (W. Zou and Yan Li 2007) proves that the histograms in high-pass bands are similar, which does

not benefit the classification. On the other hand, the histograms in low-pass bands are different. In (W. Zou and Yan Li 2007), 16 bins of histograms of projection of wavelet coefficients are exploited as an input for a neural network. Consequently, utilizing the wavelet coefficients in the low-pass band and sparse representation framework for classification would make the results more reliable as it is shown in the results (section 3.4).

The SRWC method includes two phases, the training phase and the classification phase. The training phase consists in building the dictionary from the training images and their associated classes. An overview of the training phase is given in Fig. 3.3. The classification phase allows to assign the class label of a given test image. More precisely, a feature vector is first obtained from the given test image using the same features extraction process through DWT followed by PCA as in the training phase. Its sparse codes are then computed before performing the classification based on the minimum residual criterion. The classification phase is described in sections 3.3.3 and 3.3.4 and summarized in Algorithm 3.1.

3.3.1 Single-level Discrete 2D Wavelet Transformation (DWT2)

Consider $\mathbf{I}$ as a 2D discrete-space signal (image), where $\mathbf{I}(u,v)$ denotes the pixel value. The 2D signal $\mathbf{I}(u,v)$ can be treated as 1D signals among the columns $\mathbf{I}(u,:)$ at a fixed $u-th$ row and among the rows $\mathbf{I}(:,v)$ at a fixed $v-th$ column. A single level 2D wavelet transform of an image can be captured by following the procedure in (Guo et al. 2017) using Haar kernels.

As illustrated in (Guo et al. 2017), discrete-time signals $G_L(n)$ and $G_H(n)$ are half-band low-pass and high-pass filters, respectively, defined in the spatial domain as the Haar wavelets:

$$G_H(n) = \begin{cases} 1, & 0 \leqslant n < 1/2 \\ -1, & 1/2 \leqslant n < 1 \\ 0, & otherwise \end{cases}; \quad G_L(n) = \begin{cases} 1, & 0 \leqslant n < 1 \\ 0, & otherwise \end{cases} \tag{3.1}$$

where n denotes the $n-th$ sample of the discrete-time signal.

In wavelet decomposition, the filter $G_L(n)$ is an "averaging" filter while $G_H(n)$ describes details. The 2D Discrete Wavelet Transform (DWT2) (Mallat 2008) decomposes an image into four sub-bands: average (LL), vertical (HL), horizontal (LH) and

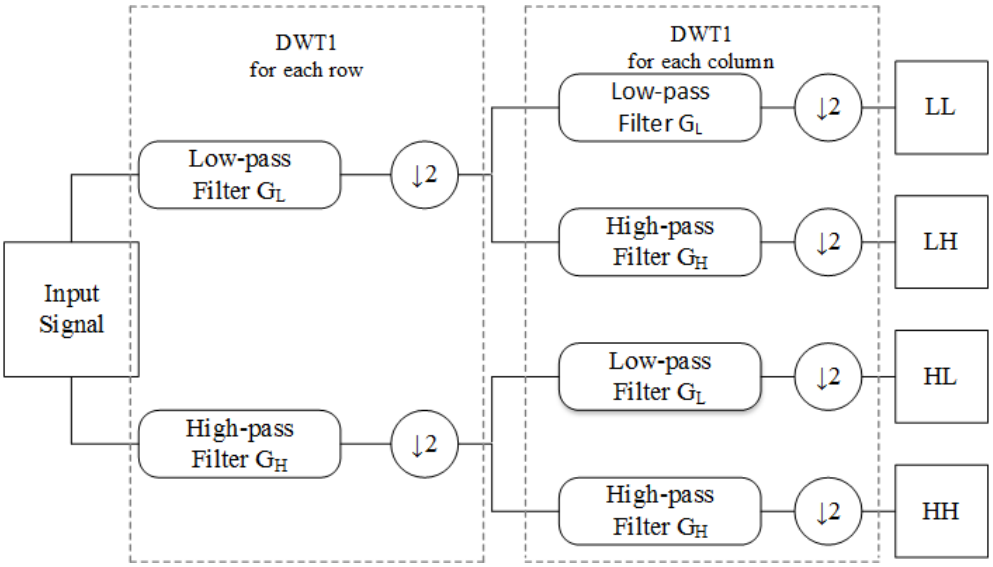


Figure 3.1: *Block chart of single-level DWT2 decomposition (Guo et al. 2017).*

diagonal (HH) information (Fig. 3.2). The details of an image, such as object's edges, are represented in the high-pass bands (LH, HL, and HH), while the primary energy of the image is represented in the low-pass band (LL). Note that after DWT2 decomposition, the combination of four sub-bands always has the same dimension as the original input image. The 2D inverse DWT (iDWT2) can trace back the DWT2 procedure by inverting the steps in Fig. 3.1. More details of wavelet transform could be found in (Mallat 2008).

3.3.2 Training phase

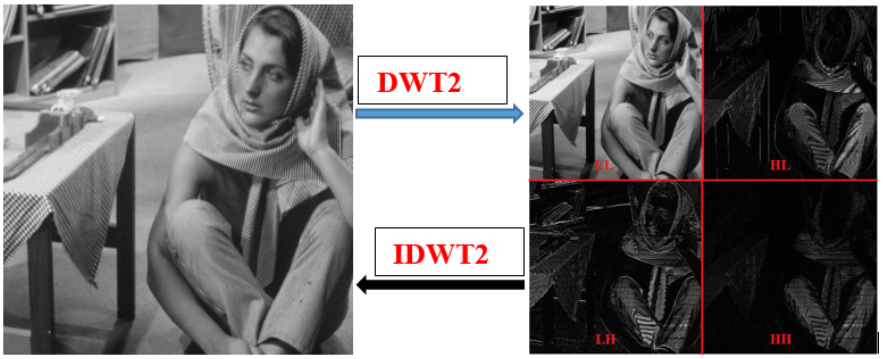
The training images are first passed through DWT2 with Haar wavelet kernel to produce 4 wavelet Sub-Bands (SB):

$$SB = \{LL, HL, LH, HH\} := DWT2(\mathbf{I}), \tag{3.2}$$

where the LL, HL, LH, and HH are sub-bands containing wavelet coefficients for average, vertical, horizontal and diagonal details of the input image.

As stated above, we will only use the LL wavelet coefficients to increase the classification accuracy. Then, principal component analysis (PCA) (I. T. Jolliffe 1986) is employed to reduce the dimension of each vectorized component, which we call an *atom*. Further, we define a new matrix $\mathbf{D}$ to describe the relations between the n atoms from all k categories ($n = \sum_{c=1...k} n_c$):

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_k] = [\mathbf{d}_{1,1}, ..., \mathbf{d}_{1,n_1}, ..., \mathbf{d}_{k,1}, ..., \mathbf{d}_{k,n_k}] \in \mathbb{R}^{m \times n} \tag{3.3}$$



(a) *Barbara image decomposed by DWT2*

(b) *Lena image decomposed by DWT2*

Figure 3.2: *Examples of image decomposed by DWT2. Left: Original image. Right: Visualization of each sub-band of wavelet coefficients.*

The training procedure to form the dictionary matrix $\mathbf{D} \in \mathbb{R}^{m \times n}$, where m is the feature dimension, is illustrated in Fig. 3.3.

3.3.3 Single test sample detection

Some discriminative models are proposed to exploit the structure of $\mathbf{D}_c$ for classification purposes (Wright, A. Y. Yang, et al. 2008). An approach is considered simple and efficient if it can model the images from a single class as lying on a linear subspace (Basri et al. 2003). Subspace models are flexible enough to capture much of the variation in real datasets. It has been observed that the images of faces under varying illuminations and expressions lie on a unique low-dimensional subspace (Basri et al. 2003). For ease of presentation, we assume that the training samples from a single class do lie on a subspace, which is only the knowledge our method will use.

Given n_c atoms of the $c-th$ category, the corresponding sub-dictionary is given by $\mathbf{D}_c = [\mathbf{d}_{c,1}, \mathbf{d}_{c,2}, ..., \mathbf{d}_{c,n_c}] \in \mathbb{R}^{m \times n_c}$. Any new feature sample $\mathbf{y} \in \mathbb{R}^m$ from the same

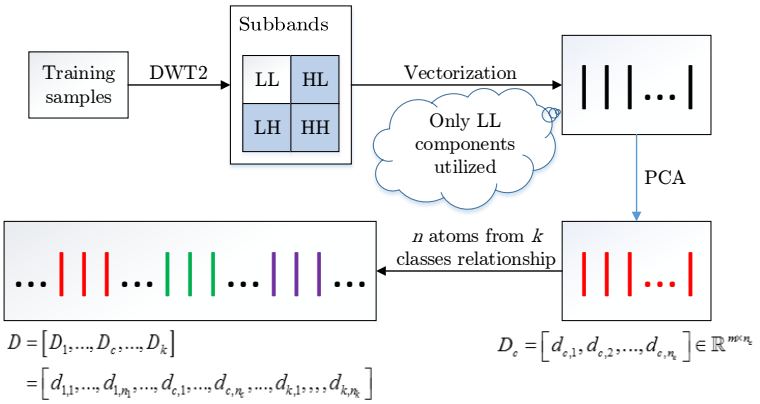


Figure 3.3: *Block chart of the dictionary training procedure of the SRWC.*

class will lie in the linear span of the atoms associated with class c as below:

$$\mathbf{y} = \mathbf{x}_{c,1}\mathbf{d}_{c,1} + \mathbf{x}_{c,2}\mathbf{d}_{c,2} + ... + \mathbf{x}_{c,n_c}\mathbf{d}_{c,n_c}, \tag{3.4}$$

where $\mathbf{x}_{c,j} \in \mathbb{R}$, $j = 1, 2, ..., n_c$. Then $\mathbf{y}$ can be rewritten as the linear combination of the entire set of atoms as below:

$$\mathbf{y} = \mathbf{D}\mathbf{x} \quad \in \mathbb{R}^m, \tag{3.5}$$

where, ideally, $\mathbf{x} = [0, ..., 0, \mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,n_c}, 0, ...0]^T \in \mathbb{R}^n$ is a sparse approximation vector whose non zero entries are those associated with the $c-th$ class. Since the entries of the vector $\mathbf{x}$ are related to the identity of the test sample $\mathbf{y}$, we are able to obtain $\mathbf{x}$ by solving the linear system of equation $\mathbf{y} = \mathbf{D}\mathbf{x}$. When in Eq. (3.5) $m < n$, the system of equations $\mathbf{y} = \mathbf{D}\mathbf{x}$ is underdetermined, and $\mathbf{x}$ cannot be found in a unique way. Further, this difficulty is resolved by taking the minimum l^2-norm solution:

$$(l_2): \quad \widehat{\mathbf{x}_2} = \arg\min \|\mathbf{x}\|_2 \quad subject\, to\, \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{3.6}$$

Note that the solution $\widehat{\mathbf{x}_2}$ from (3.6) is not instructive for recognizing the test sample $\mathbf{y}$ because $\widehat{\mathbf{x}_2}$ has a large number of nonzero entries corresponding to atoms from various classes. To resolve this difficulty in recognition, the vector $\mathbf{y}$ can be represented by only the atoms from a single class. On the other hand it is known from (Wright, A. Y. Yang, et al. 2008) that the sparser the code $\mathbf{x}$ is, the higher the accuracy of classification is. Therefore, large number of classes k is needed to make the representation of $\mathbf{y}$ sufficiently sparse to provide high accuracy of classification. This leads to the requirement to find

the sparsest solution to $\mathbf{y} = \mathbf{D}\mathbf{x}$ by solving the following optimization problem:

$$(l_0): \quad \widehat{\mathbf{x}_0} = \arg\min \|\mathbf{x}\|_0 \quad subject\,to\, \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{3.7}$$

In Eq. (3.7), $\|.\|_0$ denotes the l_0-norm, which counts the number of nonzero elements in a vector. Nevertheless, it is NP-hard to find the sparsest solution of an underdetermined system of linear equations. Fortunately, if the solution $\mathbf{x}_0$ is sparse enough, the solution of the l_0-minimization problem is equivalent to the solution to the l_1-minimization problem as follows (Wright, A. Y. Yang, et al. 2008):

$$(l_1): \quad \widehat{\mathbf{x}_1} = \arg\min \|\mathbf{x}\|_1 \quad subject\,to\, \mathbf{y} = \mathbf{D}\mathbf{x}, \tag{3.8}$$

which can be calculated in polynomial time (Elhamifar et al. 2011). Eq. 3.8 is equivalent to the Lasso problem (Tibshirani 1996) defined as follows:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{3.9}$$

In this study, we use the FISTA algorithm (Beck et al. 2009), which is an iterative method, to solve the problem in Eq. 3.8.

Fig. 3.4 illustrates the idea of SRWC that one sample is a linear combination of other samples from the same class with sparse $\mathbf{x}$.

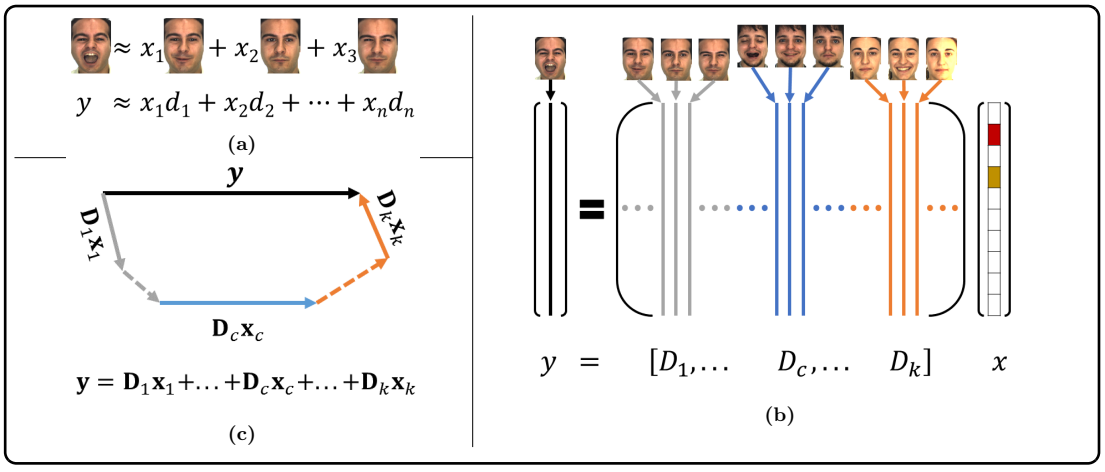


Figure 3.4: *SRWC: A sample is a linear combination of the other samples from the same class with a sparse vector* $\mathbf{x}$ *(Vu et al. 2017)*

Algorithm 3.1: Classification phase by SRWC

Input : a matrix of entire set of *atoms*, dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_k] \in \mathbb{R}^{m \times n}$ for k classes, and a test sample $\mathbf{y} \in \mathbb{R}^m$ in wavelet domain, and an error tolerance $\varepsilon > 0$

1 Normalize the columns of $\mathbf{D}$ to have unit l^2-norm.

2 Solve the l_1-minimization problem Eq. (3.8) in the form $\widehat{\mathbf{x}_1} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \, subject to \, \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leqslant \varepsilon$.

3 Compute the residuals $r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}\delta_c(\widehat{\mathbf{x}_1})\|_2 \, for \, c = 1, ..., k$

Output : $identity(\mathbf{y}) = \arg\min_c r_c(\mathbf{y})$

3.3.4 Classification phase

Given a test image, we perform the same steps of features extraction as in the training phase, using DWT2 and PCA (sections 3.3.1 and 3.3.2) to obtain the corresponding test feature vector $\mathbf{y}$. Then, we estimate its sparse representation $\widehat{\mathbf{x}_1}$ via solving the problem in Eq. (3.8). In the perfect case, the nonzero entries in the estimate $\widehat{\mathbf{x}_1}$ will be associated with the basis of the dictionary from a single class c; then we can determine the class which $\mathbf{y}$ belongs to. Nevertheless, there may be some nonzero entries associated with other categories due to noise and modeling error. To resolve this problem, $\mathbf{y}$ can be classified based on how well the coefficients in Eq. (3.4) are associated with the atoms of each object in the reconstruction of the observation $\mathbf{y}$.

For each class c, let δ_c be the characteristic function that selects the coefficients (from $\mathbf{x}$) associated only with the $c-th$ class. For $\mathbf{x} \in \mathbb{R}^n$, $\delta_c(\mathbf{x}) \in \mathbb{R}^n$ is a new vector whose only nonzero entries are the entries in $\mathbf{x}$ associated with class c. Using the nonzero entries one can approximate $\mathbf{y}$ as $\widehat{\mathbf{y}_c} = \mathbf{D}\delta_c(\widehat{\mathbf{x}_1})$, which is then classified according to a label c that minimizes the residual as follows:

$$\min_c r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}\delta_c(\widehat{\mathbf{x}_1})\|_2. \tag{3.10}$$

Algorithm 3.1 summarizes the recognition procedure.

3.4 Experimental results

The performance of the proposed SRWC is evaluated and compared with the conventional SRC (Wright, A. Y. Yang, et al. 2008), the LC-KSVD1 (Jiang et al. 2013), LC-KSVD2 (Jiang et al. 2013), and the FDDL (M. Yang, L. Zhang, Feng, et al. 2014) methods, on the three public databases, some examples of which are shown in Fig. 3.5

whose descriptions are summarized in Table 3.1. The source codes of the LC-KSVD, FDDL methods are provided by the authors of the papers (Jiang et al. 2013; M. Yang, L. Zhang, Feng, et al. 2014). As in (Elad 2010), the dimensions of the feature space extracted are sufficiently large to correctly compute the sparse representation. For SRWC, features are extracted by following the procedure stated in Section 3.3. For other methods (SRC, LC-KSVD, and FDDL), face feature descriptor is a random face, made by projecting face images onto random vectors using a random projection matrix (Wright, A. Y. Yang, et al. 2008).

3.4.1 Cross-validation

We applied Monte Carlo cross-validation (Dubitzky et al. 2007), also known as repeated random subsampling, in our experiments to better evaluate the experimental performance of the proposed SRWC method. The data is randomly separated into training and test sets in k repeated times. For each split, an image is seen in either the training set or the test set, but not in both. Then we calculate the average result over k splits. Using the Monte Carlo cross-validation helps substantially reduce the variance of the split sample error estimate and the proportion of the training-test random splits does not depend on the number k (Molinaro et al. 2005). In our experiments, we set k to 10.

3.4.2 Image databases preparation

The Extended YaleB database has 2,414 frontal-face images of 38 people (∼64 images per person) (Georghiades et al. 2001). The images are captured under various laboratory-controlled luminance states; then the images are cropped and normalized to 192x168 pixels. As suggested in (Q. Zhang et al. 2010), we randomly select 30 images from each person for training and the rest (∼34 images) for testing. After PCA, the dimension of the feature vector is $m = 650$ (Table 3.1). The number of training samples is 1140 (30 training images by 38 classes), which is also the dimension of the sparse representation vector. Some samples of this dataset are shown in Fig. 3.5a.

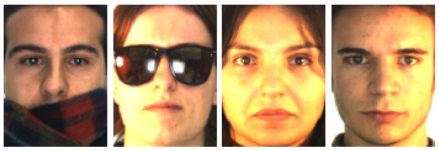
The AR face database includes over 4,000 frontal images for 126 individuals. Each subject has 26 pictures (Martinez 1998). The images are cropped to 165x120 pixels. In the experiment, a subset including 2600 images from 100 classes (50 male and 50 female) is chosen. We randomly select 20 images from every subject for training and the remaining 6 for testing. The dimension of the feature vector is $m = 900$ (Table 3.1). The total number of training samples is 2000 (20 training images by 100 classes), and it

is also the dimension of the sparse representation vector. Some samples of this dataset are shown in Fig. 3.5b.

The COIL-100 database contains 7200 color images of 100 objects captured with a black background and different lighting conditions. For every image, the authors of this dataset clipped out the object from the black background using a rectangular bounding box and resized it to 128x128 using interpolation-decimation filters to minimize aliasing. Analogously to (S. Li et al. 2016), 10 images of each object are chosen randomly for training, and the rest 62 images are used for testing in our experiment. The feature vector has the dimension $m = 1300$ (Table 3.1). The number of training samples is 1000 (10 training images by 100 classes), which is also the dimension of the sparse representation vector. A few samples of this dataset are shown in Fig. 3.5c.

All images used in our experiments are converted to the grayscale. The feature vectors obtained after PCA are all normalized to have unit norm.

(a) *Ext. YaleB (Georghiades et al. 2001)*



(b) *AR (Martinez 1998)*

(c) *COIL-100 (Nene et al. 1996)*

Figure 3.5: *Examples from three datasets.*

Table 3.1: *Description of the three datasets used in this chapter. In columns 3, 4, and 5: number of classes, number of training samples, and number of test samples, respectively.*

Database	Image size	#Class	#Training	#Test	Feature dim
Ext. YaleB	192x168	38	$n = 1140$	1274	$m = 650$
AR face	165x120	100	$n = 2000$	600	$m = 900$
COIL-100	128x128	100	$n = 1000$	6200	$m = 1300$

3.4.3 Results

Now, we present the results of the proposed SRWC method with comparison of its performance to the aforementioned methods in terms of accuracy. Moreover, in order to evaluate the robustness of the SRWC to the size of training datasets, we analyze the effect of the varying number of training samples per class.

The overall recognition rates for the Extended YaleB, AR face and COIL-100 datasets

are presented in Fig. 3.6. Our method is coded in Matlab environment and is repeated ten times for each dataset, and the average recognition rates for each method are then reported in Table 3.2. We can see that the newly proposed SRWC method outperforms the conventional SRC, and the other methods (LC-KSVD1, LC-KSVD2, FDDL) on Extended YaleB and AR, and is very close (0.13%) to the highest accuracy for COIL-100. Considering the AR (20) column in Table 3.2, one may also notice the superiority of the newly proposed SRWC over the other SRC methods. A comparison on AR (30) has not been conducted because there are not available 30 training samples per class in AR dataset (Martinez 1998).

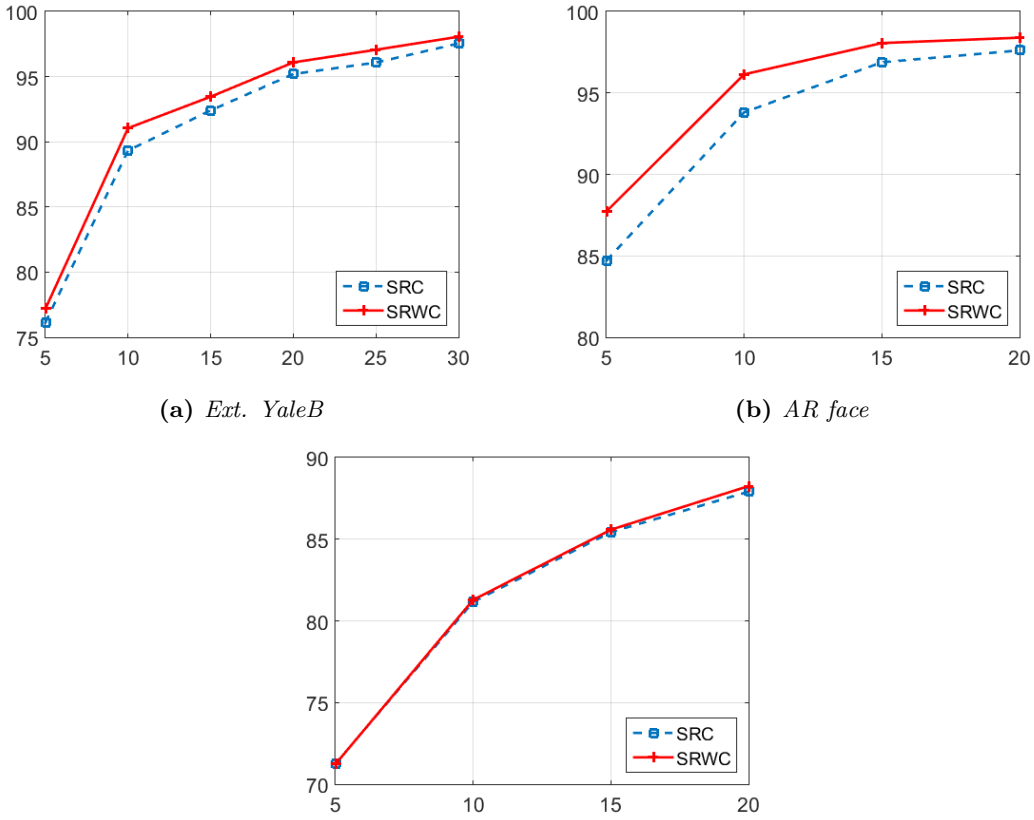
In real-world classification tasks, we often have to deal with lack of large training sets. Fig. 3.6 illustrates the accuracy of classification according to the number of training samples per class, with comparison to the conventional SRC. As it can be observed, the accuracy increases along with the gradually growing number of training samples per class and the proposed SRWC outperforms the baseline SRC (Wright, Y. Ma, et al. 2010). One may derive from Fig. 3.6 that the higher the number of atoms is used, the higher the recognition rate is. Thus, the highest recognition rate of the proposed SRWC is 98.06% for the Ext. YaleB, achieved for 30 atoms per class. Further, we determined throughout experiments that in order to receive accuracy over 80%, the number of training images should be over 10% or 15% of the size of the entire dataset.

Table 3.2: *Mean accuracy of SRWC and SRC methods. Numbers in parentheses show the training set size per class.*

	Ext. YaleB (30)	AR (20)	COIL (10)
SRC	97.54	97.61	81.16
LC-KSVD1	97.09	97.78	81.37
LC-KSVD2	97.80	97.70	**81.42**
FDDL	97.52	96.16	77.45
SRWC	**98.06**	**98.39**	81.29

3.4.4 Analysis of sparsity by visualizing the sparse representation coefficients

The performance of the proposed SRWC over the conventional SRC and the other evaluated methods can be explained by the fact that wavelets, which are naturally sparse, are used as features for sparse representation, which enhances the sparsity level of sparse codes. This advantage can be demonstrated experimentally by an analysis of the sparsity of the representation coefficients, using the visualization of i) the sum of absolute sparse codes for different test samples from a same class, ii) the sparseness



(a) *Ext. YaleB* **(b)** *AR face*

(c) *COIL-100*

Figure 3.6: *Comparison of the proposed SRWC and the SRC (Wright, A. Y. Yang, et al. 2008) method on the three evaluated databases, with classification accuracy (%) as a function of the number of training samples per class.*

measure.

i. The sum of absolute sparse codes obtained for different test samples from a same class is graphically represented with respect to the components of the sparse representation vector based on its dimension. Note that this dimension is also the number n of atoms in the dictionary ($n = k \times n_c$ where n_c is the number of training samples per class and k is the number of classes). This sum can also be represented versus classes or color bars, where each colored bar represents one class. Note that each class c is related to a set of sub-dictionary atoms. In Fig. 3.7, we present the sparsity visualization for the SRWC, using the sum of absolute sparse codes (left column) and residuals (right column) for different testing samples from the same class, on the three databases, namely Ext. Yale B (34 testing samples from 'class 33'), AR face (6 testing samples from 'class 51'), and COIL-100 (62 testing samples from 'class 69'). We can see that the class is well identified with

the highest peak or with a minimum residual.

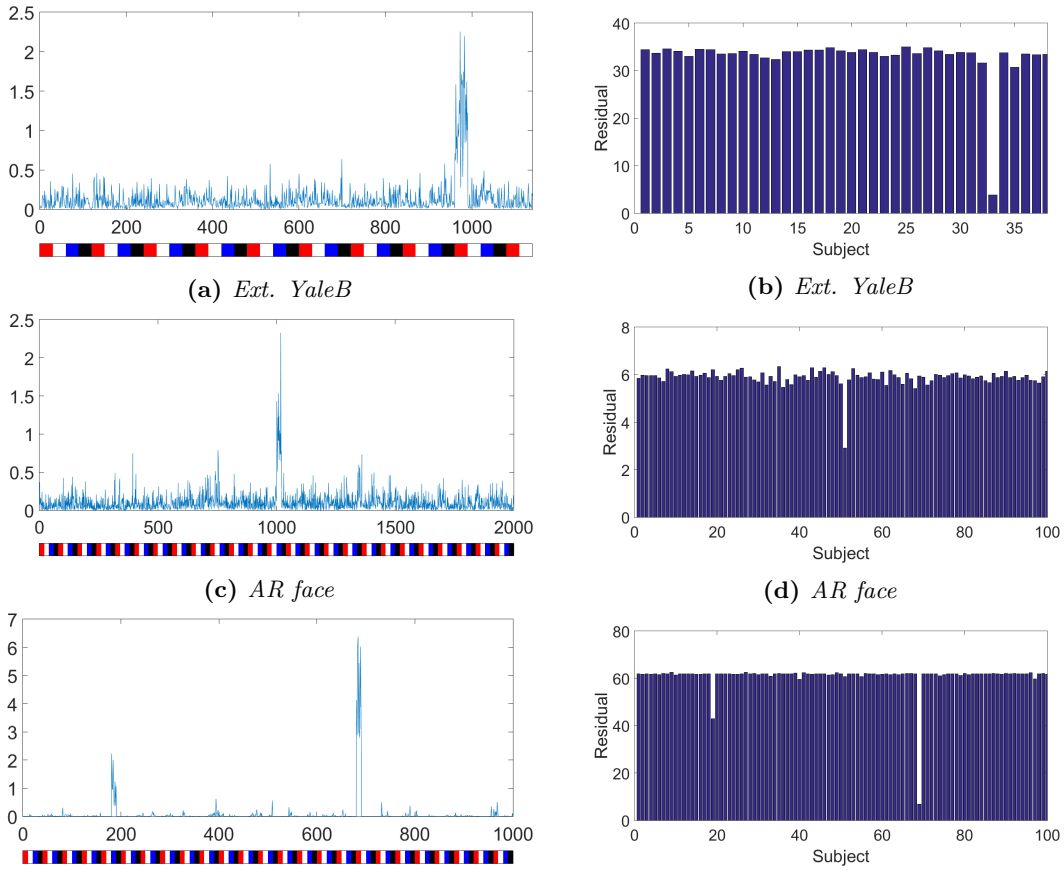
For purposes of comparison, we report in Fig. 3.8, the results of the experiments on the AR database for the SRWC and the conventional SRC method, using 6 testing samples from 'class 51'. (Indeed, as it is stated in section 3.4.2, each class of the AR database has 26 samples from which 20 ones are randomly selected for training and the remaining 6 ones for testing). The sparse representation vector has the dimension of 2000 (100 classes by 20 training samples, see Table 3.1). Hence, the X-axis denotes the dimension (2000) of the sparse representation vector, and also the classes (100) or colored bars. As a result, we can observe that the two graphs show high peaks at the $51th$ colored bar or around the $1000th - 1020th$ components of the sparse representation vector, which means that the test samples are well labeled as 'class 51'. Compared with SRC, the proposed SRWC provides better discrimination between the coefficients associated with 'class 51' and those associated with other classes. The SRWC provides the largest sparsity, which leads to the highest accuracy of classification.

ii. **The sparseness measure**: Another way to visualize the sparsity is the sparseness measure proposed in (Hoyer 2004). We adapt this concept for a sparse coefficient vector $\mathbf{x}$ as follows:

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - \|\mathbf{x}\|_1 / \|\mathbf{x}\|_2}{\sqrt{n} - 1}, \tag{3.11}$$

where n is the dimension of $\mathbf{x}$. The bigger the value of $sparseness(\mathbf{x})$ is, the sparser the vector $\mathbf{x}$ is (Hoyer 2004). To illustrate this concept, we apply Eq. 3.11 and calculate the sparseness values of the sparse codes obtained with the AR face dataset. With 600 test samples (Table 3.1), we can estimate 600 sparse codes and then calculate 600 corresponding sparseness values. These values are illustrated by the histogram in Figure 3.9. One can see that the sparseness value of the SRWC is averagely bigger than the one of the SRC (0.62 vs 0.57). More precisely, the largest sparseness value of SRWC and SRC is 0.716 and 0.665, respectively (see Figure 3.9). We may conclude from these results that SRWC provides a higher level of sparseness, which leads to better accuracy of classification.

In conclusion, the SRWC provides the higher sparsity level than the conventional

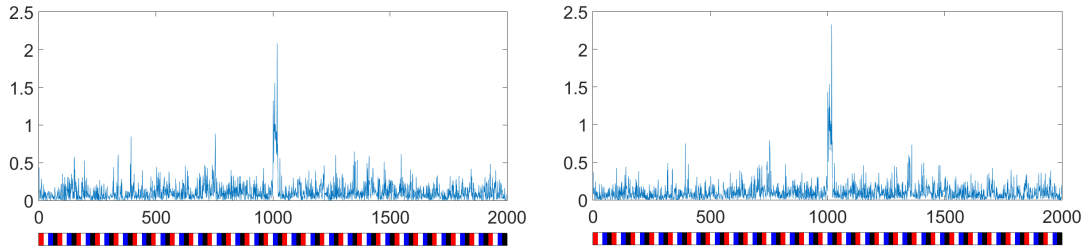


(a) *Ext. YaleB* **(b)** *Ext. YaleB*

(c) *AR face* **(d)** *AR face*

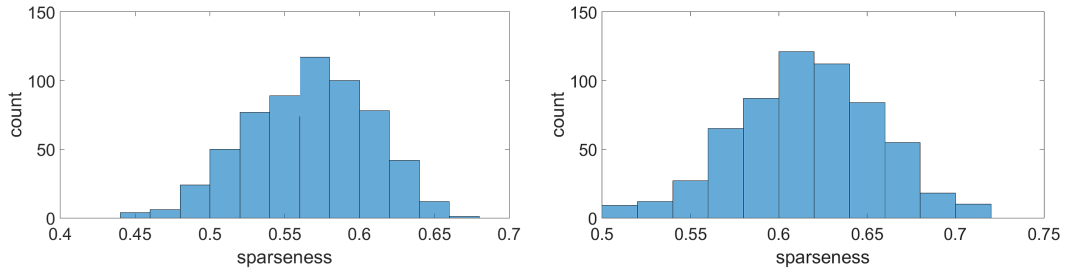
(e) *COIL-100* **(f)** *COIL-100*

Figure 3.7: *Sparsity visualization for the proposed SRWC, using the sum of absolute sparse codes (left column) and residuals (right column), for different testing samples from the same class, on the three evaluated databases, namely Ext. Yale B (34 testing samples from 'class 33'), AR face (6 testing samples from 'class 51'), and COIL-100 (62 testing samples from 'class 69'). For the graphs in left column, the X axis represents not only the dimension of sparse representation vector, but also the classes described by colored bars, from which each color denotes one class for a set of sub-dictionary atoms. On the right column, X axis simply indicates the class.*



(a) *SRC* **(b)** *SRWC*

Figure 3.8: *Sparsity visualization for the proposed SRWC and the conventional SRC (Wright, A. Y. Yang, et al. 2008), using the sum of absolute sparse codes, with 6 testing samples from 'class 51'. X axis indicates the dimension of the sparse code ($n = 2000$), and the classes described by colored bars, each color from which represents one class for a set of sub-dictionary atoms.*



(a) *SRC* **(b)** *SRWC*

Figure 3.9: *Sparseness histograms from the conventional SRC and the proposed SRWC on the AR dataset using 600 test samples.*

SRC, which leads to a better accuracy of classification.

3.5 Discussion and conclusion

The main contribution of the present study is the fusion of the low-band wavelet coefficients and the SRC approach. This fusion boosted the classification capabilities and led to an increase in the classification accuracy of the databases containing images of the same size.

The recognition rates obtained by SRWC on the two face datasets are promising, compared to LC-KSVD1 (Jiang et al. 2013), LC-KSVD2 (Jiang et al. 2013), and FDDL (M. Yang, L. Zhang, Feng, et al. 2014) (Table 3.2). Moreover, the proposed method is robust to the size of the training datasets, which is one of the advantages of the proposed method to cope with the lack of large training sets in real-world classification tasks. The SRWC demonstrated that a SRC-based method can improve its accuracy in the wavelet domain. Indeed, by taking advantage of the promoting sparsity wavelet coefficients which are used as features, and by exploiting the sparse representation of the generated features, the sparsity level of sparse codes is improved, and the proposed SRWC results in an improvement of the accuracy performance. However, the result of the SRWC on object data class (COIIL-10) is slightly lower than that of the LC-KSVC2. This can be explained by the fact that objects in such database may prone to large variation of poses, angles, or shift-variance. Indeed, DWT is not shift-invariant. It is necessary to improve the solution with a novel approach to recognise objects invariant to such variations.

In the next chapter, we extend our approach by applying Clifford Algebras, specifically Quaternion Algebras, to improve the capability of the current method.

CHAPTER 4

Contribution to the Sparse Representation Classification in the Quaternion Wavelet Domain

Chapter content

4.1 Introduction

In the previous chapter, we have introduced a SR-based approach for image classification, namely SRWC, with promising results for face and objects classifications. This chapter introduces another SR learning solution for image classification with the objective to outperform the first method, especially to further enhance the robustness of SR-based classification method (cf. Section 3.5).

Image classification is a challenge of image analysis task due to the complex nature of images with specific properties and intricate structures as well as possible noises, varying illuminations, occlusion, outliers and complex backgrounds. Hence, it is important to find a representation learning method that can capture the meaningful characteristics of the images. As it has been stated in sub-sections 1.2 and 3.1, SR is one of the most efficient and robust approaches to provide a compact representation of a signal with only a small number of meaningful features (Olshausen and Field 1997). Also, it was established that SR is the mechanism in the primary visual cortex to achieve concise description of images in terms of features (Olshausen 2003), and considered as a main principle to efficiently represent complex data (Olshausen and Field 1996). Hence, SR is an efficient representation learning method which has achieved state-of-the-art performance in signal and image processing (Elad 2010), particularly in image classification in recent years with the original sparse representation-based classification (SRC) initiated in (Wright, A. Y. Yang, et al. 2008) and its variants (Jiang et al. 2013; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010). The main idea is to estimate the sparse representation coefficients (code) of a test sample over a dictionary and then to identify its class label via the classification step based on the minimum reconstruction residual.

Although many SR-based classification methods have achieved promising performances, efforts have been necessary to improve the accuracy and to enhance the robustness of SRC methods especially for large scale systems. If existing SRC meth-

ods (Jiang et al. 2013; Wright, A. Y. Yang, et al. 2008; M. Yang, L. Zhang, Feng, et al. 2014; Q. Zhang et al. 2010) are mostly applied to the spatial domain, a recent approach for improvement of SRC, namely SRWC (Ngo et al. 2018) that we proposed and presented in Chapter 3, is performed in the sparsity-promoting Discrete Wavelet Transform (DWT) domain, making it the first SRC method in the wavelet transform domain. In this method, training and test samples are transformed into the wavelet domain and Principle Component Analysis (PCA) is used to reduce the dimension of the generated features. Then, SRWC method is performed in the wavelet domain in two steps. In the training step, the low-frequency (LL) wavelet sub-band coefficients of the training samples, after undergoing the PCA, are used to construct a dictionary. In the testing step, the features of the test samples are sparsely coded over the dictionary. Then, the class of the test samples is identified using their minimum reconstruction residuals. The advantages of DWT for classification performance have been proven in (W. Zou and Yan Li 2007). The SRWC (Chapter 3), (Ngo et al. 2018) demonstrated that SR-based classification can be performed in the Wavelet domain to enhance the sparsity level of sparse codes and discrimination ability, yielding more robust classification performance compared to the state-of-the-art SRC methods.

Despite the advantages of the wavelet domain for classification performance (W. Zou and Yan Li 2007), the DWT is restricted by its lack of shift-invariance (W. L. Chan et al. 2004). To cope with this drawback, we investigate a novel SRC method using the Quaternion Wavelet Transform (QWT) which has approximate shift-invariance and provides richer geometric information than DWT (W. L. Chan et al. 2004, 2008). Indeed, quaternion wavelet transform is a new multi-resolution image analysis tool which is based on 2D Hilbert Transform (HT), 2D analytic signal (Bulow 1999), and quaternion algebra $\mathbb{H}$. Unlike DWT whose coefficients are real, QWT is quaternion-valued, and each quaternion wavelet coefficient can be represented by amplitude and 3 phase angles, two of which encode local displacement information, the third one contains texture feature. Moreover, it can be easily computed using a dual-tree filter bank with linear computational complexity (W. L. Chan et al. 2004, 2008). Based on its interesting properties, the QWT has been applied to a number of research fields such as disparity estimation (W. L. Chan et al. 2008), image denoising (Yin et al. 2012), face recognition (YH Xu et al. 2010), texture classification (Soulard et al. 2011), and image segmentation (Subakan et al. 2011).

Motivated by the advantages of the sparsity-promoting wavelets in the AQ of the QWT and inspired by the SRWC in (Ngo et al. 2018), we propose in this chapter a novel SRC method in the Quaternion Wavelet (QW) domain to further enhance the classification performance for complex datasets. The proposed method is referred to as SRCQW which stands for Sparse Representation Classification in QW domain. To the best of our knowledge, there is no SRC approach investigated in the QW domain. In (Soulard et al. 2011), QWT is applied for image texture classification. The same transform domain is also investigated for face recognition in (YH Xu et al. 2010). In (Yi Xu et al. 2015; C. Zou et al. 2016), we have identified two main SR-based methods but they are performed in the quaternion space, in which 3 channels of color images are modelled as a quaternion. In (Yi Xu et al. 2015), a SR-based model in the quaternion is proposed for color restoration, while a SRC method is derived in the quaternion space (QSRC) for color image recognition in (C. Zou et al. 2016) (Chapter 2). Unlike these methods, the newly proposed SRC method in the QW domain benefits from the advantages of the QWT decomposition by using the QW coefficients in the low-frequency wavelet (LL) sub-bands as features to capture an efficient representation of the data with near shift-invariance property. Here, we only need features described by QW coefficients in LL sub-bands as they constitute the main component of the image. The method also benefits from the SR of the QW coefficients features to learn and capture the meaningful information of the visual data. Moreover, the construction of the dictionary and the classification are performed in the 4D space of the Algebra of Quaternions (AQ) (see Eq. 4.11). In fact, QWT decomposition on an input image yields one low-frequency QW sub-band ($\dot{LL}_q$) (approximation information) and three high-frequency QW sub-bands ($\dot{LH}_q, \dot{HL}_q, \dot{HH}_q$) (details information), where each quaternion sub-band is defined by four wavelet coefficients sub-bands. Moreover, we formulate the problem of finding the SR in the QW domain, by the novel QW Least absolute shrinkage and selection operator (QWLasso) model with quaternion l_1 minimization. To solve the QWLasso, we develop the novel Quaternion Fast Iterative Shrinkage-Thresholding (QFISTA) method, which is based on the real-valued FISTA method (Beck et al. 2009). The QFISTA maps the quaternion dictionary to a multi-dimensional real-valued matrix, composed by specific low-frequency wavelet sub-band coefficients. In addition, we develop an upper bound for the QWLasso model and use it as an approximation that establishes the iterative scheme to find the sparse coefficients of QW features. The higher separability of the

quaternion vector gives an advantage compared to the minimization model in the field of real numbers. This advantage comes from the fact that formulating the minimization model in quaternions provides additional information coming from the direction of the 4D vector, which is composed by the quaternion components. Hence, by exploiting the SR in the QW domain, with classification in the 4D space of the AQ, the proposed method improves not only the sparsity level of sparse codes for classification, but also the robustness, resulting in outperformance of the classification accuracy, as it can be seen in section 4.6.

In this method, training and test samples are transformed into the QW domain and Principle Component Analysis (PCA) is used to reduce the dimension of the generated features and computational cost. The SRCQW is performed in the QW domain in two steps. In the training step, the QW coefficients in $\dot{LL}_q$ sub-bands of the training samples, after undergoing the dimensionality reduction by PCA, are used to construct a dictionary. In the testing step, the extracted features from the test samples are sparsely coded over the dictionary. Then, the class of the test samples is identified using their minimum reconstruction residuals.

The main contributions of this method are:

i. It is based on SR learning method in the QW domain;

ii. In the training phase, a quaternion wavelet-based dictionary is constructed using four low-frequency wavelet sub-bands;

iii. In the test phase, a novel QWLasso minimization model is formulated for solving the sparse coding problem in the QW domain. In addition, the novel QFISTA method is developed in the AQ to solve the newly QWLasso model for the estimation of the sparse representation coefficients.

The experimental results show the main advantage of the proposed method, which outperforms the state-of-the-art methods including NN-based methods in terms of classification accuracy (Tables 4.2 and 4.3, where 99.6% accuracy is reported).

The rest of the chapter is organized as follows: Section 4.2 presents basic quaternion concepts; Section 4.3 describes the QWT and defines the QW coefficients of low-frequency sub-bands; Sections 4.4 develops the novel SRCQW method; Section 4.5 determines its computational complexity; Section 4.6 validates the SRCQW on commonly used datasets and compares the obtained results with several contemporary methods;

Section 4.7 concludes the chapter by discussing the contributions, the advantages, and the bottlenecks.

4.2 Basic concepts of the Algebra of Quaternions

The quaternions, which are proposed by W. R. Hamilton in 1843 (Hamilton 1844), are numbers having one real and three imaginary parts. The AQ, denoted by $\mathbb{H}$, could be regarded as a 4D Clifford algebra $Cl_{0,2}$ (Girard 2007). In this chapter, we denote scalar variables, vectors and matrices by lowercase letters (e.g., $a \in \mathbb{R}$), bold types (e.g., $\mathbf{a} \in \mathbb{R}^M$) and bold capital letters (e.g., $\mathbf{A} \in \mathbb{R}^{M \times N}$), respectively. In the AQ, a dot (above the variable) denotes quaternion variable, e.g., $\dot{a} \in \mathbb{H}$. Accordingly, vectors and matrices with quaternion entries are indicated as $\dot{\mathbf{a}} \in \mathbb{H}^M$ and $\dot{\mathbf{A}} \in \mathbb{H}^{M \times N}$, respectively.

The set of quaternions $\mathbb{H} = \{\dot{q} = q_0 + iq_1 + jq_2 + kq_3 : q_0, q_1, q_2, q_3 \in \mathbb{R}\}$ composes the AQ (Girard 2007). The three imaginary numbers satisfy the following properties:

$$i^2 = j^2 = k^2 = ijk = -1,\ ij = -ji = k,\ ik = -ki = -j,\ jk = -kj = i, \tag{4.1}$$

and its norm is: $\|\dot{q}\| = \sqrt{\dot{q}\bar{\dot{q}}} = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}$. In addition, quaternion $\dot{q}$ can also be expressed by its magnitude- phase representation as: $\dot{q} = \|\dot{q}\| e^{i\varphi} e^{j\theta} e^{k\psi}$, where $\{\varphi, \theta, \psi\}$ are the three phase angles.

Definition 4.1. *A vector with quaternion entries is called a vector of quaternions or quaternion vector:*

$$\dot{\mathbf{a}} = [\dot{a}_1, \dot{a}_2, ..., \dot{a}_M]^T \in \mathbb{H}^M, \tag{4.2}$$

where each entry is a quaternion: $\dot{a}_m = a_m^0 + a_m^1 i + a_m^2 j + a_m^3 k, m = 1, ..., M$. *Also, a quaternion vector can be formulated as:*

$$\dot{\mathbf{a}} = \mathbf{a}^0 + \mathbf{a}^1 i + \mathbf{a}^2 j + \mathbf{a}^3 k, \tag{4.3}$$

where $\mathbf{a}^e = [a_1^e, a_2^e, ..., a_M^e]^T \in \mathbb{R}^M,\ a_m^e \in \mathbb{R},\ e = 0, 1, 2, 3.$

Definition 4.2. *A matrix with quaternion entries is called a matrix of quaternions or quaternion matrix:*

$$\dot{\mathbf{A}} = [\dot{\mathbf{a}}_1, \dot{\mathbf{a}}_2, ..., \dot{\mathbf{a}}_N] \in \mathbb{H}^{M \times N}, \tag{4.4}$$

where $\dot{\mathbf{a}}_n = [\dot{a}_{1,n}, \dot{a}_{2,n}, ..., \dot{a}_{M,n}]^T \in \mathbb{H}^M$, $\dot{a}_{m,n} = a^0_{m,n} + a^1_{m,n} i + a^2_{m,n} j + a^3_{m,n} k \in \mathbb{H}$, $a^e_{m,n} \in \mathbb{R}$, $e = 0,1,2,3$, $m = 1,...,M$, $n = 1,...,N$.

A quaternion matrix can also be formulated as:

$$\dot{\mathbf{A}} = \mathbf{A}^0 + \mathbf{A}^1 i + \mathbf{A}^2 j + \mathbf{A}^3 k, \tag{4.5}$$

where $\mathbf{A}^e = [\mathbf{a}^e_1, \mathbf{a}^e_2, ..., \mathbf{a}^e_N] \in \mathbb{R}^{M \times N}, \mathbf{a}^e_n = \left[a^e_{1,n}, a^e_{2,n}, ..., a^e_{M,n}\right]^T \in \mathbb{R}^M$.

Definition 4.3. *The* $l_{1,2}$*-norm of a matrix* $\mathbf{A}$ *is:* $\|\mathbf{A}\|_{1,2} := \sum_i \|\mathbf{A}(i,:)\|_2$, *where* $\mathbf{A}(i,:)$ *denotes the* i*-th row of the matrix* $\mathbf{A}$.

The AQ is associative but non-commutative. Let $\dot{a}, \dot{b} \in \mathbb{H}$, $\lambda \in \mathbb{R}$. Based on (Girard 2007), some fundamental operations in the AQ are given as follows:

- Addition/ subtraction/ multiplication by a scalar

$$\lambda\left(\dot{a} \pm \dot{b}\right) = \lambda(a_0 \pm b_0) + \lambda(a_1 \pm b_1) i + \lambda(a_2 \pm b_2) j + \lambda(a_3 \pm b_3) k \tag{4.6}$$

- Clifford product of quaternions

$$\begin{aligned} \dot{a}\dot{b} = {} & (a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3) + (a_1 b_0 + a_0 b_1 - a_3 b_2 + a_2 b_3) i \\ & + (a_2 b_0 + a_3 b_1 + a_0 b_2 - a_1 b_3) j + (a_3 b_0 - a_2 b_1 + a_1 b_2 + a_0 b_3) k \end{aligned} \tag{4.7}$$

- Clifford product of quaternion matrices

$$\begin{aligned} \dot{\mathbf{A}}\dot{\mathbf{B}} = {} & (\mathbf{A}^0 + \mathbf{A}^1 i + \mathbf{A}^2 j + \mathbf{A}^3 k)(\mathbf{B}^0 + \mathbf{B}^1 i + \mathbf{B}^2 j + \mathbf{B}^3 k) \\ = {} & \left(\mathbf{A}^0\mathbf{B}^0 - \mathbf{A}^1\mathbf{B}^1 - \mathbf{A}^2\mathbf{B}^2 - \mathbf{A}^3\mathbf{B}^3\right) \\ & + \left(\mathbf{A}^1\mathbf{B}^0 + \mathbf{A}^0\mathbf{B}^1 - \mathbf{A}^3\mathbf{B}^2 + \mathbf{A}^2\mathbf{B}^3\right) i \\ & + \left(\mathbf{A}^2\mathbf{B}^0 + \mathbf{A}^3\mathbf{B}^1 + \mathbf{A}^0\mathbf{B}^2 - \mathbf{A}^1\mathbf{B}^3\right) j \\ & + \left(\mathbf{A}^3\mathbf{B}^0 - \mathbf{A}^2\mathbf{B}^1 + \mathbf{A}^1\mathbf{B}^2 + \mathbf{A}^0\mathbf{B}^3\right) k, \end{aligned} \tag{4.8}$$

where $\mathbf{A}^e \in \mathbb{R}^{M \times N}, \mathbf{B}^e \in \mathbb{R}^{N \times K}, e = 0,1,2,3$.

4.3 Quaternion Wavelet Transform (QWT)

The quaternion wavelets considered in this study are based on Bulow quaternion analytic signal (Bulow 1999) and a dual-tree QWT introduced in (W. L. Chan et al. 2004, 2008).

4.3.1 Quaternion Analytic Signal and Quaternion Wavelets

The quaternion analytic signal associated with a real 2D signal $s(x,y)$ is defined by its partial $(s_{\mathcal{H}_x}, s_{\mathcal{H}_y})$ and total $(s_{\mathcal{H}_{xy}})$ HTs as follows (Bulow 1999):

$$s_q(x,y) = s(x,y) + is_{\mathcal{H}_x}(x,y) + js_{\mathcal{H}_y}(x,y) + ks_{\mathcal{H}_{xy}}(x,y), \tag{4.9}$$

where $s_{\mathcal{H}_x} = s(x,y) * \frac{\delta(y)}{\pi x}, s_{\mathcal{H}_y} = s(x,y) * \frac{\delta(x)}{\pi y}, s_{\mathcal{H}_{xy}} = s(x,y) * \frac{1}{\pi^2 xy}$.
The symbol $*$ denotes the 2D convolution operation, while $\delta(x)$ and $\delta(y)$ are impulse functions along y-axis and x-axis, respectively.

In the dual-tree QWT, each quaternion wavelet is composed of four quadrature components (a real wavelet and its 2D HTs), which are organised as a quaternion: a real DWT wavelet and three imaginary wavelets obtained by 1D HT along either or both coordinates.

Denote with $\phi(t)$ and $\psi(t)$ the scaling and the wavelet functions of the 1D DWT, respectively. The 2D DWT is computed as the separable tensor products of 1D DWTs over each coordinate: the scaling function $\phi(x)\phi(y)$ and three wavelet functions $\psi(x)\psi(y)$, $\psi(x)\phi(y)$ and $\phi(x)\psi(y)$ oriented in the diagonal, vertical and horizontal directions, respectively (W. L. Chan et al. 2004, 2008).

Mathematically, the 2D QWT is defined by 1D functions as follows:

$$\psi_q^D = \psi_g(x)\psi_g(y) + i\psi_f(x)\psi_g(y) + j\psi_g(x)\psi_f(y) + k\psi_f(x)\psi_f(y) \tag{4.10a}$$

$$\psi_q^V = \psi_g(x)\phi_g(y) + i\psi_f(x)\phi_g(y) + j\psi_g(x)\phi_f(y) + k\psi_f(x)\phi_f(y) \tag{4.10b}$$

$$\psi_q^H = \phi_g(x)\psi_g(y) + i\phi_f(x)\psi_g(y) + j\phi_g(x)\psi_f(y) + k\phi_f(x)\psi_f(y) \tag{4.10c}$$

$$\phi_q = \phi_g(x)\phi_g(y) + i\phi_f(x)\phi_g(y) + j\phi_g(x)\phi_f(y) + k\phi_f(x)\phi_f(y), \tag{4.10d}$$

where $\psi_q^D, \psi_q^V, \psi_q^H$ are quaternion wavelets oriented in the diagonal, vertical and horizontal directions, respectively; the subscripts $\{g,f\}$ refer to a real-valued filter and its HT counterpart, respectively. The quaternion scaling function ϕ_q in (4.10d) corresponds to the QWT low-frequency coefficients.

4.3.2 QWT implementation

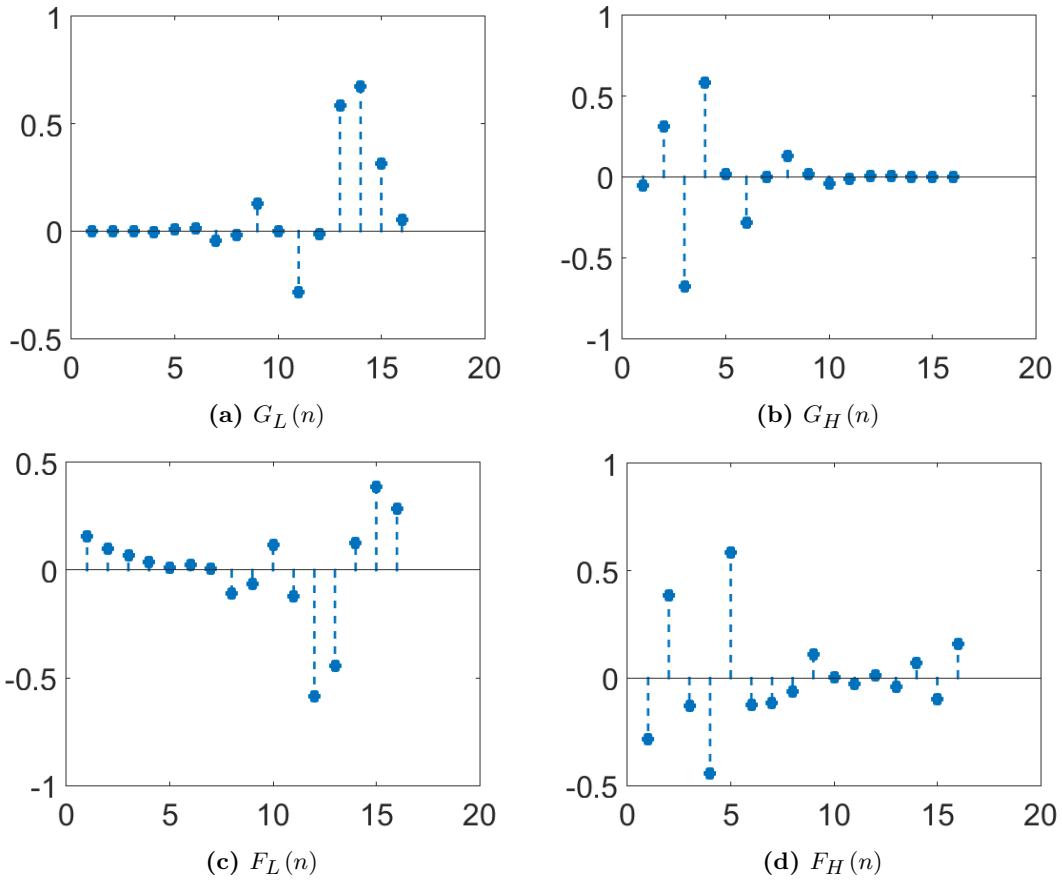


Figure 4.1: *Coefficients of the four decomposition filters are shown on the vertical axis, while the horizontal axis shows the values of n.*

The QWT is realized by using the dual-tree algorithm (W. L. Chan et al. 2008). The decomposition of an image by QWT is performed using 2D DWT, for which an input image is decomposed into a low-frequency sub-band (LL) describing the approximation information and three high-frequency sub-bands describing image details in horizontal (LH), vertical (HL) and diagonal (HH) directions, respectively (W. L. Chan et al. 2004, 2008). According to (W. L. Chan et al. 2004, 2008; Yin et al. 2012), the QWT can be implemented with the combinations of the four filters (G_L, G_H, F_L and F_H), where $G_L(n)$ and $G_H(n)$ are low-pass and high-pass wavelet filters, respectively, while $F_L(n)$ and $F_H(n)$ are the filters, corresponding to the HT of $G_L(n)$ and $G_H(n)$, respectively. Coefficients of these filters, for Daubechies 8 'db8' wavelet, are illustrated in Figure 4.1.

From these filter banks, four different wavelet coefficients are generated in the AQ to obtain the QWT coefficients. The structure of the QWT decomposition is presented in Figure 4.2, where an image is decomposed into 16 wavelet sub-bands, which construct

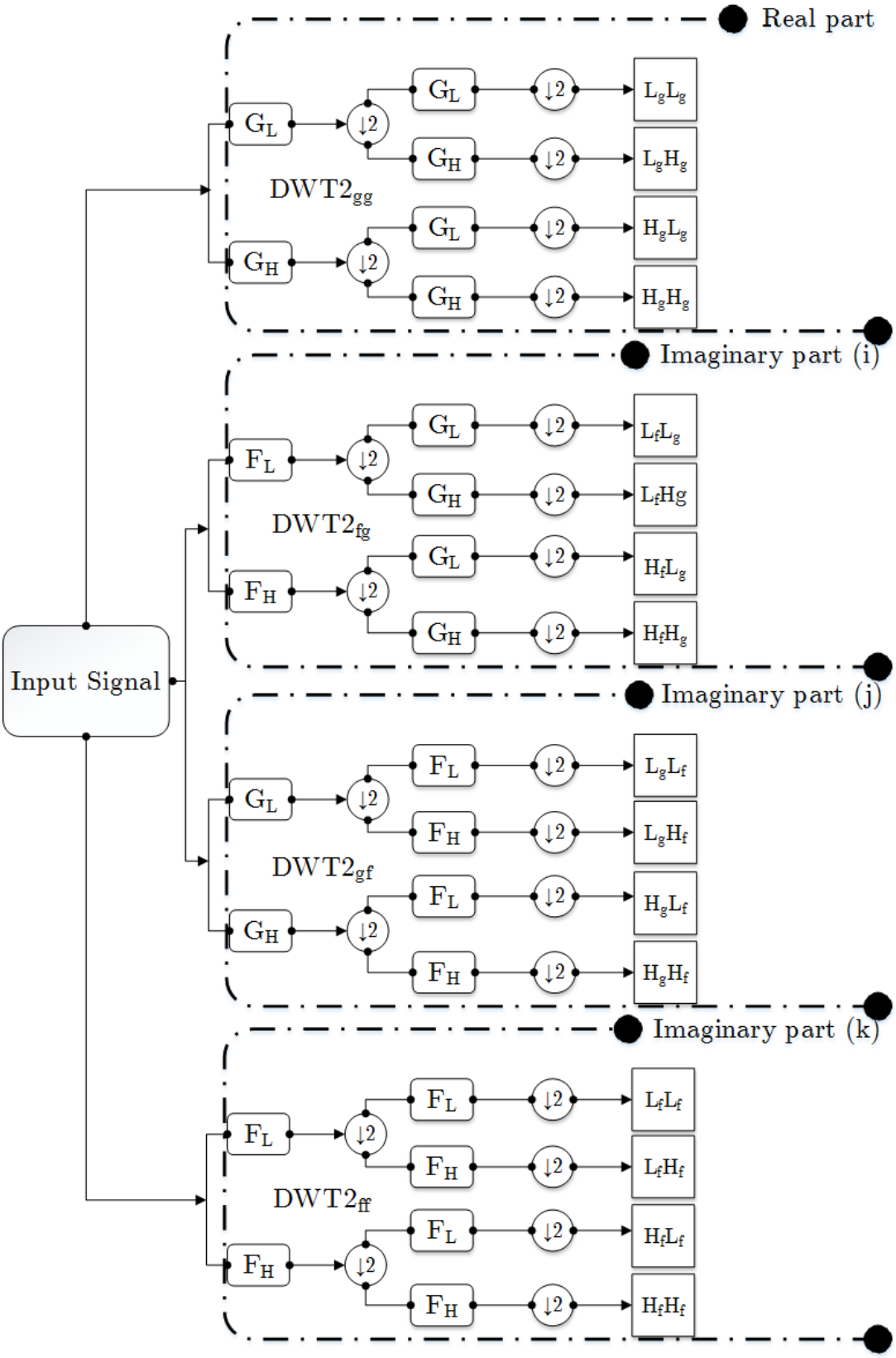
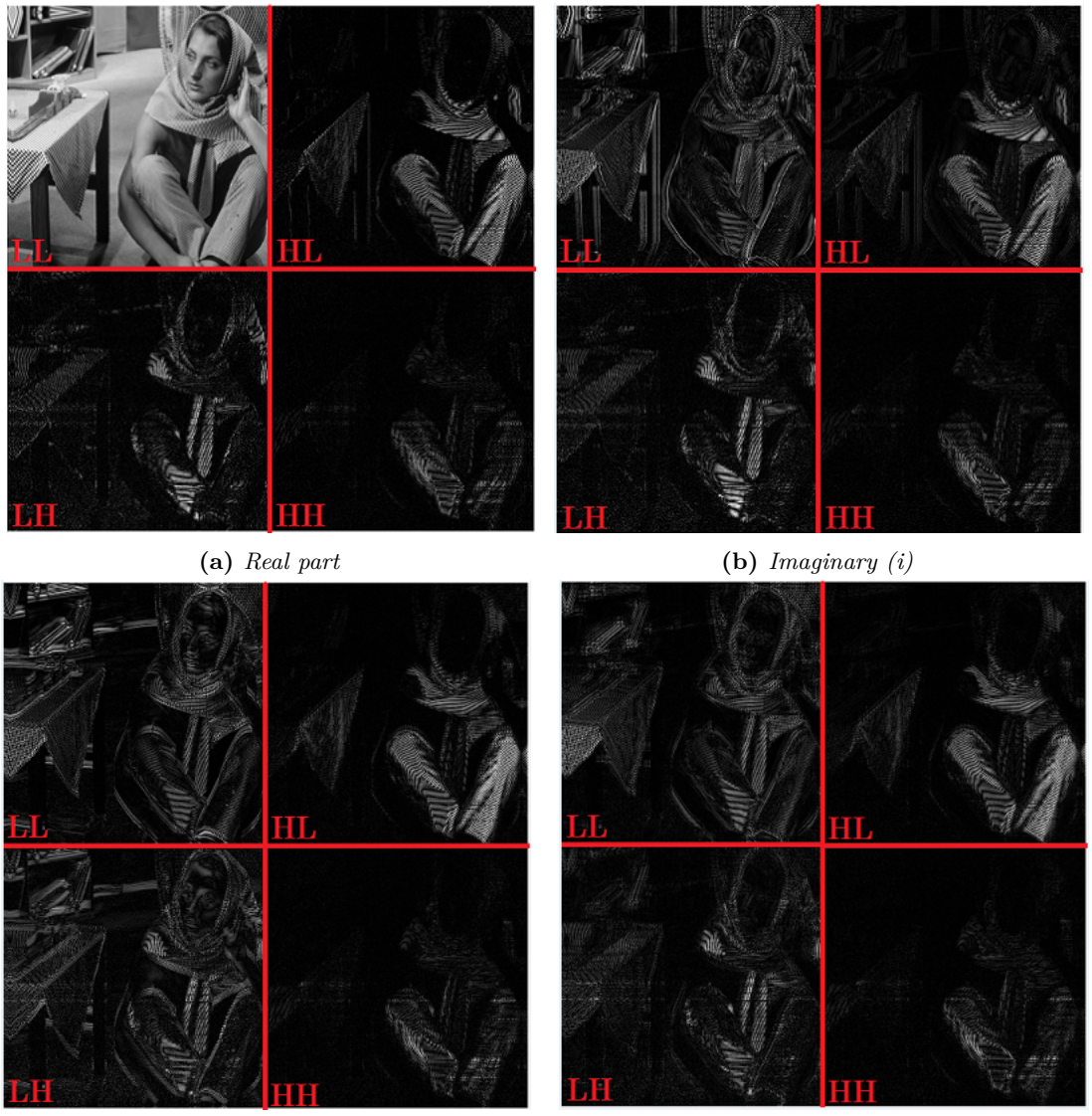


Figure 4.2: *QWT decomposition of an image.*



(a) *Real part* **(b)** *Imaginary (i)*

(c) *Imaginary (j)* **(d)** *Imaginary(k)*

Figure 4.3: *Barbara image decomposed by the QWT to 16 wavelet sub-bands as shown in Figure 4.2.* L_gL_g, L_fL_g, L_gL_f, L_fL_f *are shown at upper left side of (a), (b), (c), (d) respectively.*

four quaternion wavelet sub-bands: one low-frequency sub-band ($\dot{LL}_q$) (as in (4.11a)) and three high-frequency sub-bands ($\dot{LH}_q, \dot{HL}_q, \dot{HH}_q$) (as in (4.11b-4.11d)).

$$\dot{LL}_q = L_g L_g + iL_f L_g + jL_g L_f + kL_f L_f \tag{4.11a}$$

$$\dot{LH}_q = L_g H_g + iL_f H_g + jL_g H_f + kL_f H_f \tag{4.11b}$$

$$\dot{HL}_q = H_g L_g + iH_f L_g + jH_g L_f + kH_f L_f \tag{4.11c}$$

$$\dot{HH}_q = H_g H_g + iH_f H_g + jH_g H_f + kH_f H_f. \tag{4.11d}$$

Hence, $\dot{LL}_q$ represents the low-frequency sub-band in the QWT domain. Figure 4.3a represents the image decomposition to four real-valued wavelet sub-bands described by the real part in Figure 4.2; while the decomposition in Figures 4.3b-4.3d corresponds to the three imaginary parts of Figure 4.2.

As stated in (W. Zou and Yan Li 2007) and validated in (Ngo et al. 2018), the use of only low-frequency sub-band is sufficient to increase the accuracy of classification in the wavelet domain. We adopt this proposition and implement only the low-frequency sub-band $\dot{LL}_q$ in the further developments.

4.4 Proposed method: Sparse Representation Classification in the Quaternion Wavelet Domain (SRCQW)

In this section, we present the SRC method in the QW domain. In this method, training and test samples are transformed into the QW domain. The proposed SRCQW method is performed in two steps: dictionary construction and classification in the QW domain. The dictionary is constructed from the training samples and mapped into the 4D space of AQ. In the classification step, assuming a given dictionary, we first estimate the quaternion SR of a query image by solving the novel QWLasso model applying the novel QFISTA method (quaternion sparse coding stage in section 4.4.2.2); then we compute the class-dependent residual to identify the label of the query image (namely label assignment stage in section 4.4.2.3). An overview of the novel SRCQW method is illustrated in Figure 4.4, while the main procedure is summarized in Algorithm 4.1.

4.4.1 Training phase and dictionary of low-frequency sub-bands of the QWT

In this chapter, a quaternion vector, which belongs to the basis of a quaternion dictionary (matrix whose entries are quaternions composed of low-frequency sub-bands), is called a

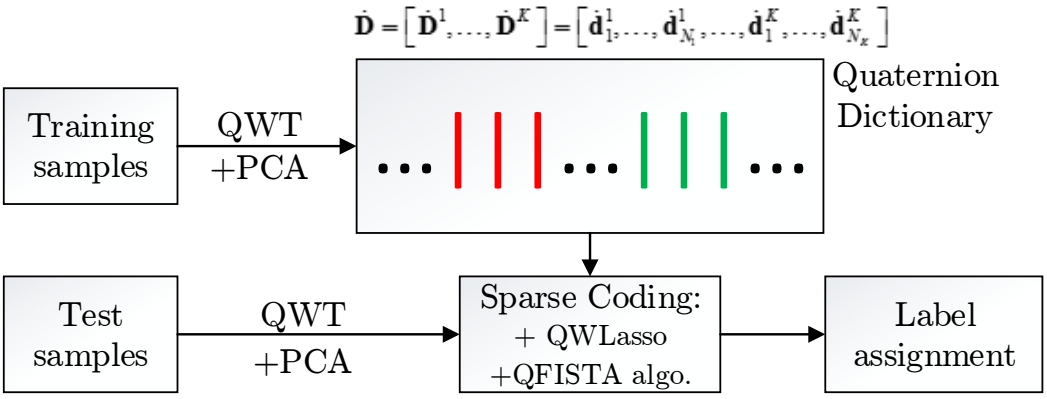


Figure 4.4: *The framework of the proposed SRCQW method.*

quaternion atom. Consider a classification problem with K classes. Let N_k be the number of training samples from the *k-th* class for $1 \leqslant k \leqslant K$. The QWT is first performed on every training sample, as shown in Figure 4.2, to obtain the low-frequency QW sub-band (given with Eq.4.11a) as image descriptor (atom). Next, in the *k-th* class, N_k atoms are arranged as vector columns of quaternions, generating hence a dictionary on which PCA (I. Jolliffe 2011) is applied to reduce the dimension of the quaternion atoms by compressing them onto a lower-dimensional feature space with dimension M, yielding thus the dictionay for the class k: $\dot{\mathbf{D}}^k = \left[\dot{\mathbf{d}}_1^k, \dot{\mathbf{d}}_2^k, ..., \dot{\mathbf{d}}_{n_k}^k, ..., \dot{\mathbf{d}}_{N_k}^k\right] \in \mathbb{H}^{M \times N_k}$, where the quaternion atoms $\dot{\mathbf{d}}_{n_k}^k \in \mathbb{H}^M, n_k = 1, ..., N_k$. Now, using the training dictionaries $\dot{\mathbf{D}}^k, k = 1, ..., K$, we compose the dictionary for all classes:

$$\dot{\mathbf{D}} = \left[\dot{\mathbf{D}}^1, \dot{\mathbf{D}}^2, ..., \dot{\mathbf{D}}^K\right] = \left[\dot{\mathbf{d}}_1^1, ..., \dot{\mathbf{d}}_{N_1}^1, \dot{\mathbf{d}}_1^2, ..., \dot{\mathbf{d}}_{N_2}^2, ..., \dot{\mathbf{d}}_1^K, ..., \dot{\mathbf{d}}_{N_K}^K\right]. \tag{4.12}$$

The total number of atoms in the dictionary $\dot{\mathbf{D}}$ is $N = \sum_{k=1...K} N_k$. Employing Eq. 4.5, we present the dictionary $\dot{\mathbf{D}}$ as follow:

$$\dot{\mathbf{D}} = \mathbf{D}^0 + \mathbf{D}^1 i + \mathbf{D}^2 j + \mathbf{D}^3 k \quad \in \mathbb{H}^{M \times N}, \tag{4.13}$$

where $\mathbf{D}^e = \left[\mathbf{D}^{e,1}, \mathbf{D}^{e,2}, ..., \mathbf{D}^{e,K}\right] = [\mathbf{d}_1^{e,1}, ..., \mathbf{d}_{N_1}^{e,1}, \mathbf{d}_1^{e,2}, ..., \mathbf{d}_{N_2}^{e,2}, ..., \mathbf{d}_1^{e,K}, ..., \mathbf{d}_{N_K}^{e,K}] \in \mathbb{R}^{M \times N}$, $e = 0, 1, 2, 3$. Each training dictionary $\mathbf{D}^e$ presents information from a single low-frequency sub-band only. For example, $\mathbf{D}^0$ is constructed by the $L_g L_g$ from Eq. 4.11a. It follows from Eq. 4.3 that a quaternion atom can be represented as a quaternion vector

of real-value vectors:

$$\dot{\mathbf{d}}_{n_k}^{k} = \mathbf{d}_{n_k}^{0,k} + \mathbf{d}_{n_k}^{1,k} i + \mathbf{d}_{n_k}^{2,k} j + \mathbf{d}_{n_k}^{3,k} k,\ \mathbf{d}_{n_k}^{e,k} \in \mathbb{R}^M,\ e = 0,1,2,3,\ k = 1,...,K. \tag{4.14}$$

By comparing the coefficients from Eq. 4.14 with those in Eq. 4.11a, it is clear that for the n_k-th training image from class k, the entities $\mathbf{d}_{n_k}^{0,k}$, $\mathbf{d}_{n_k}^{1,k}$, $\mathbf{d}_{n_k}^{2,k}$, and $\mathbf{d}_{n_k}^{3,k}$ are respective vectors of the low-frequency sub-band for real part $(L_g L_g)$, imaginary parts (i) $L_f L_g$, (j) $L_g L_f$, and (k) $L_f L_f$. An image example for each vector is presented with every upper left image in Figure 4.3.

4.4.2 Classification in the QW domain

In this section, finding the SR of a test sample is necessary to identify its class label. Within this scope, we formulate the QWLasso model in the QW domain and propose the QFISTA method to resolve this problem. Then a classifier minimizing a residual criterion is used to identify the membership of the image.

Algorithm 4.1: Classification phase by SRCQW

Input : a test quaternion vector $\dot{\mathbf{y}} \in \mathbb{H}^M$, the quaternion dictionary matrix $\dot{\mathbf{D}} \in \mathbb{H}^{M \times N}$ for K classes in quaternion wavelet domain and the parameter λ from Eq. 4.16.

1 Normalize the columns of $\dot{\mathbf{D}}$ to have unit l_2-norm.

2 Compute the quaternion sparse vector via the QWLasso model:
$\widehat{\dot{\mathbf{x}}} = \underset{\dot{\mathbf{x}} \in \mathbb{H}^N}{\arg\min} \left\{ \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\dot{\mathbf{x}} \right\|_2^2 + \lambda \|\dot{\mathbf{x}}\|_1 \right\}$ (Go to Algorithm 4.2).

3 Compute the residuals: $r_k(\dot{\mathbf{y}}) = \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\delta_k\left(\widehat{\dot{\mathbf{x}}}\right) \right\|_2$ $for\ k = 1,...,K.$

Output : $identity(\dot{\mathbf{y}}) = \underset{k}{\arg\min}\, r_k(\dot{\mathbf{y}})$.

4.4.2.1 Sparse Representation Quaternion Wavelet model

In (C. Zou et al. 2016), the authors extended the SRC and developed the QLasso model to work in the AQ, where every quaternion represents the three color channels, at every image pixel, and preserves the correlation information among the channels. In contrast, the proposed SRCQW method uses four low-frequency sub-bands in the AQ. In (C. Zou et al. 2016), to solve the QLasso model and calculate the sparse vector, the authors applied the ADMM approach, while in the present study, we develop the novel optimization algorithm, QFISTA, which allows for solving the QWLasso model defined with Eq. 4.16. Another difference between the QLasso (C. Zou et al. 2016) and the

QWLasso, used by the proposed SRCQW method, is that the former is formulated in 3D sub-space of the 4D space of the AQ, while the latter is formulated in the entire 4D space of the AQ. The higher dimension is expected to provide higher accuracy of classification.

Assume unknown test image. By applying QWT we obtain the low-frequency quaternion $\dot{LL}_q$. After the reduction of its dimension with the PCA and following Eq.4.14 we determine its form as a quaternion of approximation information $\dot{\mathbf{y}} = \mathbf{y}^0 + \mathbf{y}^1 i + \mathbf{y}^2 j + \mathbf{y}^3 k \ \in \mathbb{H}^M$, where $\mathbf{y}^e \in \mathbb{R}^M,\ e = 0,1,2,3$. Accordingly, the quaternion SR of $\dot{\mathbf{y}}$ is $\dot{\mathbf{x}} = \mathbf{x}^0 + \mathbf{x}^1 i + \mathbf{x}^2 j + \mathbf{x}^3 k, \dot{\mathbf{x}} \in \mathbb{H}^N,\ \mathbf{x}^e \in \mathbb{R}^N$. Then, we develop the SR quaternion wavelet model to estimate vector $\hat{\dot{\mathbf{x}}}$ of the quaternion sparse vector $\dot{\mathbf{x}}$:

$$\hat{\dot{\mathbf{x}}} = \underset{\dot{\mathbf{x}} \in \mathbb{H}^N}{\arg\min} \|\dot{\mathbf{x}}\|_1, \quad s.t.\ \dot{\mathbf{y}} = \dot{\mathbf{D}}\dot{\mathbf{x}}, \tag{4.15}$$

where $\dot{\mathbf{D}} \in \mathbb{H}^{M \times N}$ is the entire quaternion dictionary built up by the N quaternion atoms. The symbol $\|\dot{\mathbf{x}}\|_1 := \sum_i |\dot{x}_i|$ denotes the l_1-norm of the quaternion vector, which is naturally based on the l_1-norm of the real-valued vector. Inspired from (Wright, A. Y. Yang, et al. 2008), we use Eq. 4.15 to formulate the new QWLasso model as follows:

$$\hat{\dot{\mathbf{x}}} = \underset{\dot{\mathbf{x}} \in \mathbb{H}^N}{\arg\min} \left\{ \left\| \dot{\mathbf{y}} - \dot{\mathbf{D}}\dot{\mathbf{x}} \right\|_2^2 + \lambda \|\dot{\mathbf{x}}\|_1 \right\}, \tag{4.16}$$

where $\lambda > 0$ is the regularization parameter, which controls the sparsity of $\dot{\mathbf{x}}$ and provides a trade-off between the sparsity penalty and the fidelity term. To solve Eq. 4.16, we propose a novel gradient-based algorithm called QFISTA that we introduce in the next sub-section..

4.4.2.2 Quaternion sparse coding stage

Inspired by the results obtained with the FISTA method (Beck et al. 2009), we develop its version to operate in the AQ to resolve Eq. 4.16.

Paper (C. Zou et al. 2016) demonstrated that calculations in quaternions are considerably more complex than those in real number system. This greatly increases the complexity of solving the optimization problem in Eq. 4.16 with the novel SRCQW method. Therefore, to simplify the solution of Eq. 4.16 (also point 2 in Algorithm 4.1), we map the quaternion dictionary of low-frequency sub-bands to a real-valued dictionary implemented with Algorithm 4.2. To develop the mapping, we apply Eqs. 4.3 and 4.5

and rewrite equation $\dot{\mathbf{y}} = \dot{\mathbf{D}}\dot{\mathbf{x}}$ in the form:

$$\mathbf{y}^0 + \mathbf{y}^1 i + \mathbf{y}^2 j + \mathbf{y}^3 k = (\mathbf{D}^0 + \mathbf{D}^1 i + \mathbf{D}^2 j + \mathbf{D}^3 k)(\mathbf{x}^0 + \mathbf{x}^1 i + \mathbf{x}^2 j + \mathbf{x}^3 k). \tag{4.17}$$

Then, using Eq. 4.7, the above Eq. 4.17 can be rewritten as:

$$\begin{aligned} \mathbf{y}^0 + \mathbf{y}^1 i + \mathbf{y}^2 j + \mathbf{y}^3 k = {} & \mathbf{D}^0\mathbf{x}^0 - \mathbf{D}^1\mathbf{x}^1 - \mathbf{D}^2\mathbf{x}^2 - \mathbf{D}^3\mathbf{x}^3 \\ & + \left(\mathbf{D}^1\mathbf{x}^0 + \mathbf{D}^0\mathbf{x}^1 - \mathbf{D}^3\mathbf{x}^2 + \mathbf{D}^2\mathbf{x}^3\right) i \\ & + \left(\mathbf{D}^2\mathbf{x}^0 + \mathbf{D}^3\mathbf{x}^1 + \mathbf{D}^0\mathbf{x}^2 - \mathbf{D}^1\mathbf{x}^3\right) j \\ & + \left(\mathbf{D}^3\mathbf{x}^0 - \mathbf{D}^2\mathbf{x}^1 + \mathbf{D}^1\mathbf{x}^2 + \mathbf{D}^0\mathbf{x}^3\right) k \end{aligned} \tag{4.18}$$

By comparing the coefficients of i, j and k in both sides of Eq. 4.18 we obtain:

$$\begin{bmatrix} \mathbf{y}^0 \\ \mathbf{y}^1 \\ \mathbf{y}^2 \\ \mathbf{y}^3 \end{bmatrix} = \begin{bmatrix} \mathbf{D}^0 & -\mathbf{D}^1 & -\mathbf{D}^2 & -\mathbf{D}^3 \\ \mathbf{D}^1 & \mathbf{D}^0 & -\mathbf{D}^3 & \mathbf{D}^2 \\ \mathbf{D}^2 & \mathbf{D}^3 & \mathbf{D}^0 & -\mathbf{D}^1 \\ \mathbf{D}^3 & -\mathbf{D}^2 & \mathbf{D}^1 & \mathbf{D}^0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^0 \\ \mathbf{x}^1 \\ \mathbf{x}^2 \\ \mathbf{x}^3 \end{bmatrix}. \tag{4.19}$$

Since $\mathbf{D}^e$ $(e = 0,1,2,3)$ are real-valued matrices we solve Eq. 4.19 in the field of real numbers. To develop the solution to Eq. 4.16, we use the operators $\mathbf{P}$ and $\mathbf{Q}$, which naturally arise from Eq. 4.19. Consider $\dot{\mathbf{D}} = \mathbf{D}^0 + \mathbf{D}^1 i + \mathbf{D}^2 j + \mathbf{D}^3 k \in \mathbb{H}^{M\times N}$, $\mathbf{D}^e \in \mathbb{R}^{M\times N}$, which is a quaternion matrix. There exists a unique operator $\mathbf{P} : \mathbb{H}^{M\times N} \to \mathbb{R}^{4M\times 4N}$:

$$\mathbf{P}\left(\dot{\mathbf{D}}\right) := \begin{bmatrix} \mathbf{D}^0 & -\mathbf{D}^1 & -\mathbf{D}^2 & -\mathbf{D}^3 \\ \mathbf{D}^1 & \mathbf{D}^0 & -\mathbf{D}^3 & \mathbf{D}^2 \\ \mathbf{D}^2 & \mathbf{D}^3 & \mathbf{D}^0 & -\mathbf{D}^1 \\ \mathbf{D}^3 & -\mathbf{D}^2 & \mathbf{D}^1 & \mathbf{D}^0 \end{bmatrix} \in \mathbb{R}^{4M\times 4N}. \tag{4.20}$$

The existence and uniqueness of $\mathbf{P}$ follow from Eqs. 4.17-4.19. Accordingly, for any quaternion vector $\dot{\mathbf{x}} = \mathbf{x}^0 + \mathbf{x}^1 i + \mathbf{x}^2 j + \mathbf{x}^3 k, \dot{\mathbf{x}} \in \mathbb{H}^N$, $x^e \in \mathbb{R}^N$, $e = 0,1,2,3$, we define the operator $\mathbf{Q} : \mathbb{H}^N \to \mathbb{R}^{4N}$ such that:

$$\mathbf{Q}(\dot{\mathbf{x}}) := \left[(\mathbf{x}^0)^T (\mathbf{x}^1)^T (\mathbf{x}^2)^T (\mathbf{x}^3)^T\right]^T \in \mathbb{R}^{4N}. \tag{4.21}$$

Denote with $\mathbf{Q}^{-1}$ the inverse of $\mathbf{Q}$, i.e. $\mathbf{Q}^{-1}(\mathbf{Q}(\dot{\mathbf{x}})) := \dot{\mathbf{x}}$. As proven in (C. Zou

Algorithm 4.2: QFISTA with constant step-size

Input : The test quaternion vector $\dot{\mathbf{y}} \in \mathbb{H}^M$, the quaternion dictionary $\dot{\mathbf{D}} \in \mathbb{H}^{M\times N}$ and regularization parameter $\lambda \in \mathbb{R}$.

1 Initialization: $\mathbf{u} := Q(\dot{\mathbf{y}}), \mathbf{B} := P\left(\dot{\mathbf{D}}\right), \mathbf{s} := Q(\dot{\mathbf{x}}), \mathbf{s}_0 = 0, \mathbf{w}_1 = \mathbf{s}_0 \in \mathbb{R}^M$, $t_1 = 1$.

2 Step r. $(r \geqslant 1)$ Compute:

i. $\mathbf{s}_r = p_L(\mathbf{w}_r)$ where $p_L(\mathbf{w}_r) = \arg\min_{\mathbf{s}_r}\{h(\mathbf{s}_r)\} = \arg\min_{\mathbf{s}_r}\{g(\mathbf{s}_r) + \varphi(\mathbf{s}_r)\}$ (Beck et al. 2009) and it is computed by Algorithm 4.3, which determines the current $\mathbf{s}_r$,

ii. $t_{r+1} = \frac{1+\sqrt{1+4t_r^2}}{2}$, coefficient used in $\mathbf{w}_{r+1}$ update,

iii. $\mathbf{w}_{r+1} = \mathbf{s}_r + \frac{t_r - 1}{t_{r+1}}(\mathbf{s}_r - \mathbf{s}_{r-1})$.

Output : $\widehat{\dot{\mathbf{x}}} = \mathbf{Q}^{-1}(\mathbf{s}_r)$.

et al. 2016), the operators $\mathbf{P}$ and $\mathbf{Q}$ possess the following properties:

(i) $\mathbf{P}$ and $\mathbf{Q}$ are linear;

(ii) $\|\mathbf{Q}(\dot{\mathbf{x}})\|_2 = \|\dot{\mathbf{x}}\|_2, \quad \forall \dot{\mathbf{x}} \in \mathbb{H}^N$;

(iii) $\mathbf{Q}\left(\dot{\mathbf{D}}\dot{\mathbf{x}}\right) = \mathbf{P}(\dot{\mathbf{D}})\mathbf{Q}(\dot{\mathbf{x}}), \ \forall \dot{\mathbf{D}} \in \mathbb{H}^{M\times N}, \forall \dot{\mathbf{x}} \in \mathbb{H}^N$;

(iv) $\left\|\dot{\mathbf{D}}\dot{\mathbf{x}}\right\|_2 = \left\|\mathbf{P}(\dot{\mathbf{D}})\mathbf{Q}(\dot{\mathbf{x}})\right\|_2, \ \forall \dot{\mathbf{D}} \in \mathbb{H}^{M\times N}, \forall \dot{\mathbf{x}} \in \mathbb{H}^N$.

Further, we define the operator $qmat(\mathbf{Q}(\dot{\mathbf{x}})) : \mathbb{R}^{4N\times 1} \to \mathbb{R}^{N\times 4}$ (C. Zou et al. 2016):

$$qmat(\mathbf{Q}(\dot{\mathbf{x}})) := \left[\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\right] \quad \in \mathbb{R}^{N\times 4}. \tag{4.22}$$

Now we apply the operators $\mathbf{P}$, $\mathbf{Q}$ and $qmat$ on the QWLasso model in Eq. 4.16 and map it from the AQ to the field of real numbers:

$$\mathbf{Q}\left(\widehat{\dot{\mathbf{x}}}\right) = \underset{Q(\dot{\mathbf{x}})\in\mathbb{R}^{4N}}{\arg\min}\left\{\left\|\mathbf{Q}(\dot{\mathbf{y}}) - \mathbf{P}\left(\dot{\mathbf{D}}\right)\mathbf{Q}(\dot{\mathbf{x}})\right\|_2^2 + \lambda\|qmat(\mathbf{Q}(\dot{\mathbf{x}}))\|_{1,2}\right\} \tag{4.23}$$

Note that in Eq. 4.23 we apply the $l_{1,2}$-norm because the quaternion $\dot{\mathbf{x}}$ in the l_1 term in Eq. 4.16 is mapped to a matrix in Eq. 4.23. For the sake of simplicity, we denote $\mathbf{u} := \mathbf{Q}(\dot{\mathbf{y}}) \in \mathbb{R}^{4M}, \mathbf{B} := \mathbf{P}\left(\dot{\mathbf{D}}\right) \in \mathbb{R}^{4M\times 4N}$, $\mathbf{s} := \mathbf{Q}(\dot{\mathbf{x}}) \in \mathbb{R}^{4N}$, and rewrite Eq. 4.23 in the following concise form:

$$\widehat{\mathbf{s}} = \underset{\mathbf{s}\in\mathbb{R}^{4N}}{\arg\min}\left\{\|\mathbf{u} - \mathbf{B}\mathbf{s}\|_2^2 + \lambda\|qmat(\mathbf{s})\|_{1,2}\right\} \tag{4.24}$$

Algorithm 4.3: The Optimal Gradient Method for Computing '**s**'

Input : Lipschitz constant $L, \mathbf{w}_r, \mathbf{s}_{r,1} = 0$, $maxIter$ - the maximum number of iterations, and the parameter $\lambda \in \mathbb{R}$.

1 Calculate $\nabla f(\mathbf{w}_r)$ using Eq. 4.26.

2 Calculate $h_1^{best} = h(\mathbf{s}_{r,1})$, assign $\mathbf{s}_{r,1}^{best} = \mathbf{s}_{r,1}$.

3 **While** $1 \leqslant z \leqslant maxIter$ compute

i. $\nabla h(\mathbf{s}_{r,z})$, apply Eq. 4.27,

ii. $\mathbf{s}_{r,z+1} = \mathbf{s}_{r,z} - \frac{\nabla h(\mathbf{s}_{r,z})}{\sqrt{z}\|\nabla h(\mathbf{s}_{r,z})\|_2}$,

iii. $h_{z+1}^{best} = \min\left\{h_z^{best}, h(\mathbf{s}_{r,z+1})\right\}$,

iv. $\mathbf{s}_{r,z+1}^{best} = \begin{cases} \mathbf{s}_{r,z}^{best}, & if\ h_{z+1}^{best} = h_z^{best} \\ \mathbf{s}_{r,z+1}, & if\ h_{z+1}^{best} = h(s_{r,z+1}) \end{cases}$,

end while

Output : $\mathbf{s}_r = \mathbf{s}_{r,z+1}^{best}$.

Then, we split Eq. 4.24 into two parts and denote them: $f(\mathbf{s}) = \|\mathbf{u} - \mathbf{B}\mathbf{s}\|_2^2$ and $g(\mathbf{s}) = \lambda\|qmat(\mathbf{s})\|_{1,2}$. Hence, Eq. 4.24 becomes: $\widehat{\mathbf{s}} = \underset{\mathbf{s}\in\mathbb{R}^{4N}}{\arg\min}\{f(\mathbf{s}) + g(\mathbf{s})\}$. Note that $f(\mathbf{s})$ is smooth since its second derivative exists. It follows that in the vicinity of $\mathbf{s}, \exists\ \mathbf{w} \in \mathbb{R}^{4N}$ and a Lipschitz constant L, which define a quadratic form $\varphi(\mathbf{s})$ that approximates $f(\mathbf{s})$:

$$\begin{aligned} f(\mathbf{s}) &\leqslant f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{s} - \mathbf{w}) + \frac{L}{2}\|\mathbf{s} - \mathbf{w}\|_2^2 \\ &\approx \frac{L}{2}\left\|\mathbf{s} - \left(\mathbf{w} - \frac{1}{L}\nabla f(\mathbf{w})\right)\right\|_2^2 = \varphi(\mathbf{s}). \end{aligned} \tag{4.25}$$

Unlike the original FISTA, which uses l_1-norm, we implement Eq. 4.24 with the $l_{1,2}$-norm. The proposed QFISTA is applied to approximate the solution of Eq. 4.16 with the help of Eq. 4.25. In order to implement the novel QFISTA and calculate the sparse code $\widehat{\mathbf{x}}$, we develop the novel Algorithm 4.2 and its sub-procedure Algorithm 4.3 to solve equation $p_L(\mathbf{w}) = \underset{\mathbf{s}}{\arg\min}\{h(\mathbf{s})\}$ for $\mathbf{s}$, where $h(\mathbf{s}) = g(\mathbf{s}) + \varphi(\mathbf{s})$.

To apply Algorithms 4.2 and 4.3, the formulas given below are calculated, where

$\mathbf{w}, L$, and $\nabla f(\mathbf{w})$ are real-valued constants:

$$\nabla f(\mathbf{s}) = -2\mathbf{B}^T(\mathbf{u} - \mathbf{B}\mathbf{s}) = 2\left(\mathbf{B}^T\mathbf{B}\mathbf{s} - \mathbf{B}^T\mathbf{u}\right), \tag{4.26}$$

$$\nabla h(\mathbf{s}) = \nabla g(\mathbf{s}) + \nabla\varphi(\mathbf{s}), where \tag{4.27}$$

$$\nabla g(\mathbf{s}) = \lambda\left[\begin{array}{cccc} \frac{\partial g(\mathbf{s})}{\partial \mathbf{s}_1} & \frac{\partial g(\mathbf{s})}{\partial \mathbf{s}_2} & \cdots & \frac{\partial g(\mathbf{s})}{\partial \mathbf{s}_{4N}} \end{array}\right], \nabla\varphi(\mathbf{s}) = L\left(\mathbf{s} - \left(\mathbf{w} - \frac{1}{L}\nabla f(\mathbf{w})\right)\right). \tag{4.28}$$

4.4.2.3 Label assignment stage

In the final stage, we compute the class dependent residual for each class k, $1 \leqslant k \leqslant K$ as: $r_k(\dot{\mathbf{y}}) = \left\|\dot{\mathbf{y}} - \dot{\mathbf{D}}\delta_k\left(\widehat{\dot{\mathbf{x}}}\right)\right\|_2$, where δ_k denotes the characteristic function that selects the coefficients from $\widehat{\dot{\mathbf{x}}}$ associated with class k. For $\widehat{\dot{\mathbf{x}}} \in \mathbb{H}^N$, the non-zeros entries of $\delta_k\left(\widehat{\dot{\mathbf{x}}}\right) \in \mathbb{H}^N$ are associated with the class k. Finally, the test quaternion vector $\dot{\mathbf{y}}$ is assigned to the class providing the minimal residual.

4.5 Computational complexity

The computational complexity for the SRCQW algorithm is estimated as the number of arithmetic operations required to calculate $\mathbf{s}_{k,z+1}^{best}$ in Algorithms 4.2 and 4.3 for solving Eq. 4.16. For the purpose of simplicity, we assume that i) the number of training samples (number of dictionary atoms) is the same for every class and equals $n = \max\{N_k\}_{k=1}^{K}$. ii) each iterative algorithm requires same q iterations to converge.

We observe, from Eqs. 4.26 and 4.27, that the most computationally expensive expressions are: $\mathbf{B}^T\mathbf{B}\mathbf{s} - \mathbf{B}^T\mathbf{u}$ and $\nabla g(\mathbf{s}) + \nabla\varphi(\mathbf{s})$, where $\mathbf{B}^T\mathbf{u}$ can be precomputed. Note $\mathbf{B} \in \mathbb{R}^{4M\times 4N}, \mathbf{u} \in \mathbb{R}^{4M}$, where N denotes the total number of atoms in the quaternion dictionary. We obtain that $\mathbf{B}^T\mathbf{u} = O(NM)$. Since $N = Kn$, where K is the number of classes, we have $\mathbf{B}^T\mathbf{u} = O(nKM)$. Consider that $\mathbf{s} \in \mathbb{R}^{4N}$. There are two different ways to calculate the computational complexity of the chain: $\mathbf{B}^T(\mathbf{B}\mathbf{s}) = \left(\mathbf{B}^T\mathbf{B}\right)\mathbf{s}$. For the left one we have $\mathbf{B}^T(\mathbf{B}\mathbf{s}) = O(MN + MN) = O(MN)$ number of arithmetic operations. Concerning the right chain of matrix multiplication we have $\left(\mathbf{B}^T\mathbf{B}\right)\mathbf{s} = O(NMN + N^2) = O(N^2M)$. Therefore, we select $O(NM) = O(nKM)$ for the computational complexity of $\mathbf{B}^T\mathbf{B}\mathbf{s}$.

Consider Eqs. 4.27-4.28. It is straightforward to show that $\nabla g(\mathbf{s}) = O(nK)$ and $\nabla\varphi(\mathbf{s}) = O(nK)$. Now, consider that each of the Algorithms 4.2 and 4.3 requires q iterations to converge. Therefore, following Eqs. 4.24, 4.26, 4.27, we evaluate the computational complexity of the SRCQW method as:

$$\underbrace{O(nKM)}_{\mathbf{B}^T\mathbf{u}} + q[\underbrace{O(nKM)}_{\mathbf{B}^T\mathbf{Bs}} + q[\underbrace{O(nK)}_{\nabla g(\mathbf{s})} + \underbrace{O(nK)}_{\nabla \varphi(\mathbf{s})}]]$$
$$= O(nKM) + O(qnKM) + O\left(q^2nK\right) \approx O(qnKM) + O\left(q^2nK\right).$$

The last big-O expression implies that the SRCQW method has a computational complexity that equals to the $\max\{O(qnKM), O(q^2nK)\}$.

4.6 Experimental results

In order to validate the SRCQW capabilities, we carried out experiments on four public datasets and compare its results with several contemporary methods in the field including SR-based methods, namely SRC (Wright, A. Y. Yang, et al. 2008), LC-KSVD (Jiang et al. 2013), FDDL (M. Yang, L. Zhang, Feng, et al. 2014), LRSDL (Vu et al. 2017), and SRWC (Ngo et al. 2018) as well as NNs, namely Centralined PLN (Liang et al. 2018), Distributed PLN (Liang et al. 2018), PCANet1 (T.-H. Chan et al. 2015), and Deep/Wide Net (Alom et al. 2018).

4.6.1 Cross-validation

To evaluate the performance of the proposed SRCQW method, Monte Carlo cross-validation (Dubitzky et al. 2007) was used. It randomly splits the dataset into a training set and a test set and repeats this process k times. For each split, a sample appears in either the training set or the test set, but not in both. The results are then averaged over the k splits. The advantage of using the Monte Carlo cross-validation is that it can substantially reduce the variance of the split sample error estimate and the proportion of the training-test random splits does not depend on the number k (Molinaro et al. 2005). In our experiments, k is set to 10.

4.6.2 Details of datasets

The validation public databases are: the Extended YaleB face dataset (Georghiades et al. 2001), the AR face dataset (Martinez 1998), the AR gender dataset (Martinez 1998) and a multi-class object category dataset – the COIL-100 (Nene et al. 1996). Figure 4.5 shows examples from these datasets (Georghiades et al. 2001; Martinez 1998; Nene et al. 1996), whose descriptions are summarized in Table 4.1.

i. The Extended YaleB dataset (Georghiades et al. 2001) contains face images of 38

people. For every face, about 64 images are taken under various conditions. For each face, we randomly select 30 images for training, making a total number of 1140 training and 1274 test images.

ii. The AR face dataset (Martinez 1998) contains frontal faces of 126 people and 26 images are available for each face. In this experiment, we use 100 people (50 males and 50 females). For every person, 20 images are randomly selected for training and 6 for testing.

iii. The AR gender dataset (Martinez 1998) is generated by choosing 14 non-occluded images per individual from 100 people (50 males and 50 females), having total of 1400 images. We randomly select 350 images from each class (male/female) for training; the remaining images are used for testing.

iv. The COIL-100 dataset (Nene et al. 1996) consists of 7200 color images of 100 objects. For every experiment, we randomly select 10 images of each object for training and the rest (62 images) for testing.

Table 4.1: *Description of the four datasets used in this chapter. In columns 3, 4, and 5: number of classes, number of training samples, and number of test samples, respectively.*

Database	Image size	#Class	#Training	#Test	Feature dim
Ext. YaleB	192x168	38	$N = 1140$	1274	$M = 2400$
AR face	165x120	100	$N = 2000$	600	$M = 2200$
AR gender	165x120	100	$N = 700$	700	$M = 1700$
COIL-100	128x128	100	$N = 1000$	6200	$M = 400$

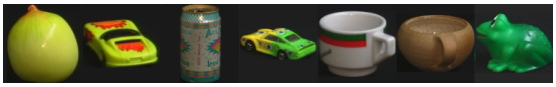
4.6.3 Effect of varying parameter λ on the classification accuracy

Figure 4.6 shows the performance of the proposed SRCQW on three datasets using different values of the regularization parameter λ (Eq.4.16, Algorithm 4.2) from the discrete set $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$ (C. Zou et al. 2016). We observe that the highest accuracy is obtained with $\lambda = 10^{-3}$ for the face datasets and with $\lambda = 10^{-2}$ for the COIL-100. Hereafter, we utilize these λ values.

4.6.4 Overall classification accuracy

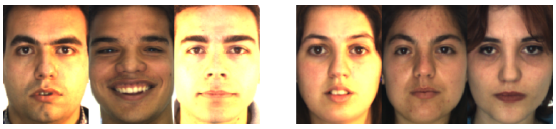
To evaluate the SRCQW classification performance, we compare it with five contemporary SR-based methods on the four datasets. The chosen methods comprise a fundamental one (Wright, A. Y. Yang, et al. 2008), its elaborations (Jiang et al. 2013;

(a) *Ext. YaleB (Georghiades et al. 2001)*



(b) *COIL-100 (Nene et al. 1996)*

(c) *AR face (Martinez 1998)*



(d) *AR gender (Martinez 1998)*

Figure 4.5: *Examples from the four datasets.*

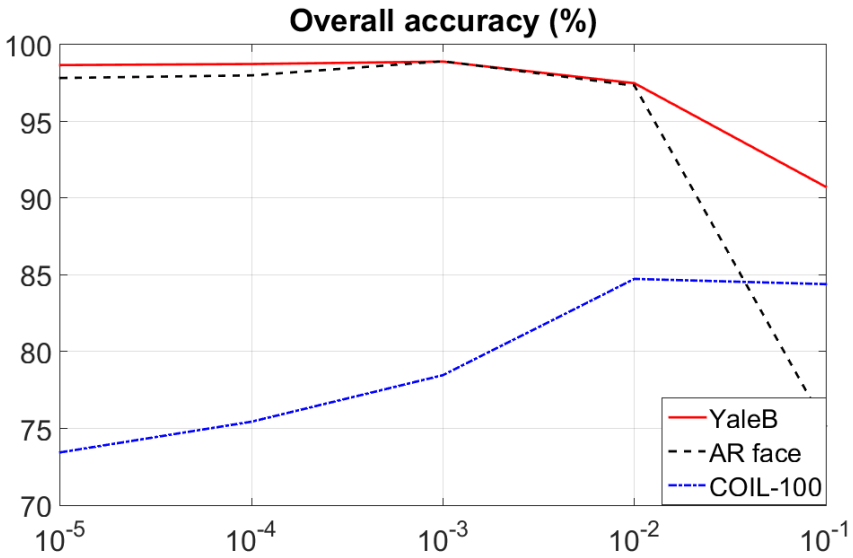


Figure 4.6: *Correlation between λ (x axis), in Eq.4.16 and Algorithm 4.2, and accuracy.*

Table 4.2: *Mean accuracy (%) of the proposed SRCQW and contemporary SR-based methods. Numbers in parentheses show the training set size per class. Best results are marked in bold.*

	SRC	LC-KSVD1	LC-KSVD2	FDDL	LRSDL	SRWC	SRCQW
Ext. YaleB (30)	97.5	97.1	97.8	97.5	98.8	98.1	**98.9**
Ext. YaleB (55)	99	N/A	N/A	98.77	99.3	98.7	**99.85**
AR (20)	97.6	97.8	97.7	96.2	98.8	98.4	**98.9**
ARgender (350)	92.6	88.4	90.1	93.7	95.4	96.5	**98.6**
COIL-100 (10)	81.2	81.4	81.4	77.5	84.4	81.3	**84.8**

Vu et al. 2017; M. Yang, L. Zhang, Feng, et al. 2014), and a SRC in the wavelet domain (Ngo et al. 2018). For each dataset, we execute the code ten times in Matlab environment using same number of training samples for all runs, but randomly select different test set for every run. The average classification rates are reported in Table 4.2. It is evident that the proposed SRCQW outperforms the competitors in all cases. In particular, a substantial improvement, up to 2.1%, has been made for the ARgender dataset, while in the classification of the Ext. YaleB database, the SRCQW achieved the very high 99.85%. For the object dataset COIL-100, except for the LRSDL method, the new SRCQW improves the classification accuracy with up to 3.4%. Compared to the LC-KSVC1 and LC-KSVC2, the proposed SRCQW has superior performance for object dataset (COIL-100). Indeed, as it has been noticed for the SRWC method in chapter 3, which has slightly lower performance than LC-KSVD2 (also in Table 4.2), the proposed SRC in the QW domain shows better performance on object dataset (COIL-100), thus demonstrating the benefits of the shift-invariance property of QWT.

Table 4.3: *Comparision of the novel SRCQW with recent Neural Networks. In (T.-H. Chan et al. 2015; Liang et al. 2018), results are not reported on COIL-100 database. It is the same case for the two face datasets in (Alom et al. 2018).*

	Ext. YaleB			AR face			COIL-100		
	#Training	**#Test**	**Acc**	**#Training**	**#Test**	**Acc**	**#Training**	**#Test**	**Acc**
Centralined PLN (Liang et al. 2018)	1600	800	96.6	1800	800	95.3	N/A	N/A	N/A
Distributed PLN (Liang et al. 2018)	1600	800	96.8	1800	800	95.6	N/A	N/A	N/A
PCANet1 (T.-H. Chan et al. 2015)	N/A	N/A	97.8	N/A	N/A	98	N/A	N/A	N/A
Deep Net (Alom et al. 2018)	N/A	N/A	N/A	N/A	N/A	N/A	5000	2200	94.1
Wide Net (Alom et al. 2018)	N/A	N/A	N/A	N/A	N/A	N/A	5000	2200	**96.8**
SRCQW	1596	818	**99.6**	1800	800	**98.5**	5000	2200	95.6
SRCQW	1140	1274	98.9	1500	1100	98	N/A	N/A	N/A

To further reveal the advantage of the proposed SRCQW, we compare it with five recent contemporary NNs approaches. In (Liang et al. 2018), the authors trained a large deep architecture with a progressive learning network in a distributed setup and used the ADMM optimizing algorithm. PCANet1 (T.-H. Chan et al. 2015) is a deep learning network, which is constructed on cascaded PCA, binary hashing, and blockwise histograms. The authors employed PCA to learn multistage filter banks. Then binary

hashing and block histograms for indexing and pooling were implemented. In (Alom et al. 2018), deep and wide convolutional NNs (CNNs) are implemented within the Energy Efficient Deep Neuromorphic Networks framework released by IBM in 2016. Table 4.3 reveals that the proposed SRCQW outperforms the five NNs on the two face datasets despite using less training images, as it can be seen in the last row from Table 4.3 (1140 instead of 1600 with Ext. YaleB and 1500 instead of 1800 with AR face). When using nearly the same number of training images, as it is depicted in the penultimate row from Table 4.3, SRCQW achieves very high accuracies of 99.6% on the YaleB and 98.5% on the AR classification. With COIL-100 dataset, SRCQW obtains very high result (95.6%) with such complicated dataset and ranks 2nd in the table.

4.6.5 Accuracy versus feature dimensions

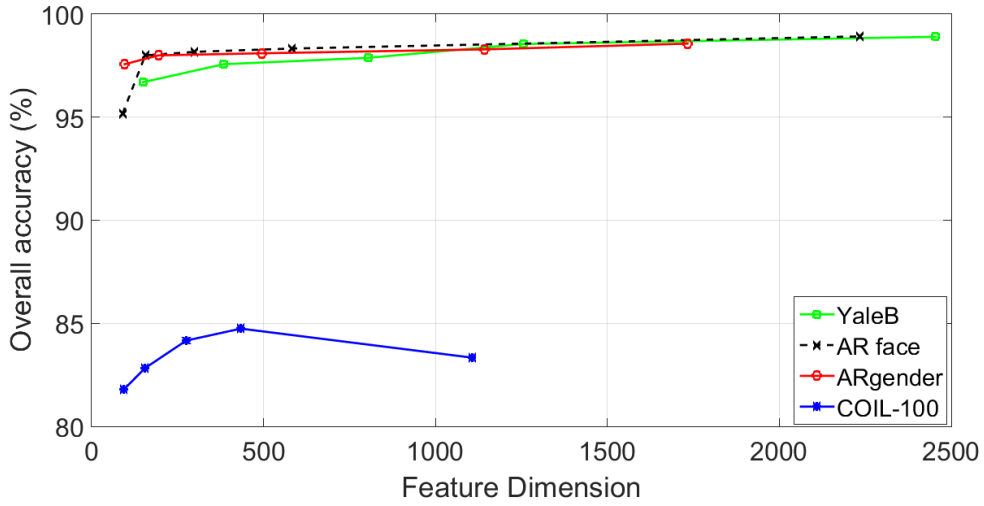
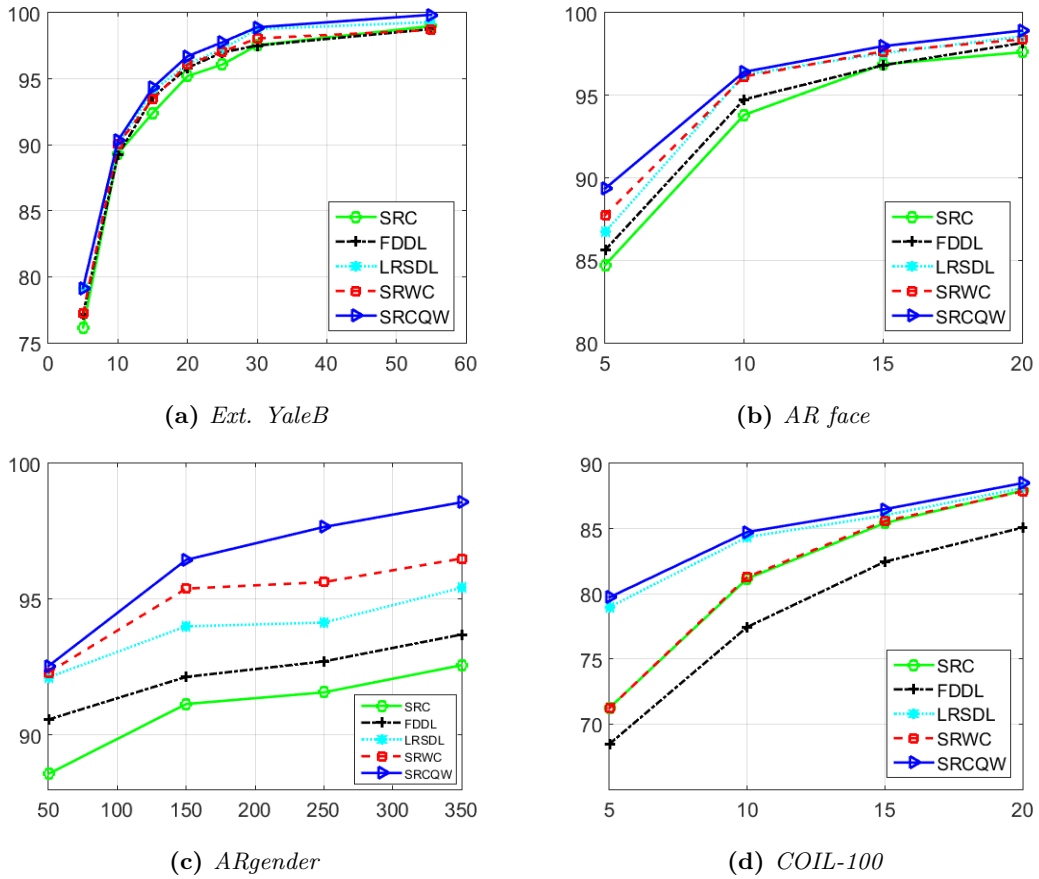


Figure 4.7: *Effect of various feature dimensions (shown on X axis) on the overall accuracy.*

In order to study the classification accuracy of the novel SRCQW method and its sensitivity to the dimension (M) of feature vectors (obtained after PCA for the reduction of its dimension), we conducted experiments with varying dimensions M of the atoms used in the dictionary. For this purpose, we used the four baseline datasets and illustrated the results in Figure 4.7. The maximum overall accuracies on the Ext. YaleB, AR face, AR gender, and COIL-100 are 98.9%, 98.92%, 98.57%, and 84.74% respectively, obtained for dimensions 2400, 2200, 1700, and 400 respectively. These results tell that if the dimension of the atom increases, the accuracy of face recognition increases as well, while in case of COIL-100, the accuracy of objects recognition decreases when the dimension increase from 400 to 1100.

4.6.6 Accuracy versus size of training set



(a) *Ext. YaleB* **(b)** *AR face*

(c) *ARgender* **(d)** *COIL-100*

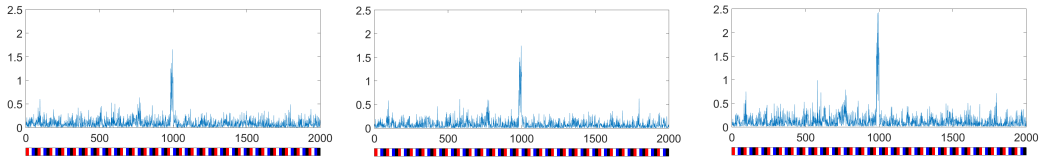
Figure 4.8: *Comparison of SRCQW with main competitors on the four datasets, with overall classification accuracy (%) as a function of the number of training samples per class.*

To further show the SRCQW capabilities, we investigate its robustness according to the size of the training set. Hence, we conducted additional experiments where we vary the number of training samples per class. For the purpose of comparison, we use again the datasets: Ext. YaleB, AR face, AR gender and COIL-100. Figure 4.8 shows that the accuracy curve of SRCQW is above the curves of the other methods. Moreover, the novel SRCQW method achieved an accuracy of 90% on the YaleB, 96% on the AR face, and about 85% on the COIL-100 databases using 10 training samples. One may derive from Fig. 4.8 that the higher the number of atoms is used, the higher the recognition rate is. Thus, the highest recognition rate of the proposed SRCQW is 99.85% for the Ext. YaleB, achieved for 55 atoms per class. Further, we concluded throughout experiments that to receive accuracy over 80%, we need to use at least 10% of the total images of the dataset for training.

4.6.7 Analysis of sparsity by visualizing the sparse representation coefficients

Analogous to the work in section 3.4.4, to experimentally demonstrate the advantage of quaternion wavelet over classical wavelet and real domain, we conducted an analysis of the sparsity of the representation coefficients obtained from the proposed SRCQW, SRWC, and SRC methods, using the visualization of i) the sum of absolute sparse codes for different test samples from a same class, ii) the sparseness measure.

i. **The sum of absolute sparse codes**: One testing class from the AR face dataset is used for illustration. For this purpose, we calculate the sum of sparse codes of six test samples from 'class 50' in the AR face dataset. As described in Table 4.1, 2000 samples are used for training. Thus, each sparse code has the dimension of 2000 entries as illustrated along the X axis in Figure 4.9. Each colored rectangle from the horizontal bars represents one class (1 to 100 classes) for a subset of dictionary atoms. One can see that the three graphs possess high peaks at the $50th$ colored rectangle or around the $1000th$ component of the quaternion SR vector, which means that the test samples are well labeled as 'class 50'. By comparison with SRC and SRWC, the proposed SRCQW provides better discrimination between the coefficients associated with 'class 50' and those associated with other classes.



(a) *SRC* **(b)** *SRWC* **(c)** *SRCQW*

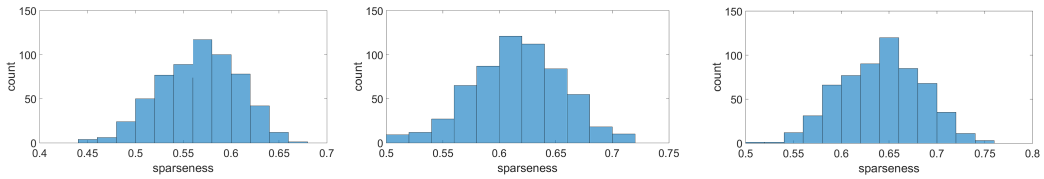
Figure 4.9: *Sparse codes generated by the SRCQW, SRC, and SRWC methods. X axis shows the sparse codes dimensions, Y axis shows the sum of sparse codes for different test samples.*

ii. **The sparseness measure**: Another way to visualize the sparsity is the sparseness measure proposed in (Hoyer 2004). We adapt this concept for a quaternion sparse vector $\dot{\mathbf{x}}$ as follows:

$$sparseness(\dot{\mathbf{x}}) = \frac{\sqrt{N} - \|\dot{\mathbf{x}}\|_1 / \|\dot{\mathbf{x}}\|_2}{\sqrt{N} - 1}, \tag{4.29}$$

where N is the dimension of $\dot{\mathbf{x}}$. The bigger $sparseness(\dot{\mathbf{x}})$ is, the sparser $\dot{\mathbf{x}}$ is (Hoyer 2004). To illustrate this concept, we apply Eq. 4.29 and calculate the sparseness values of the sparse codes obtained with the AR face dataset. With 600

test samples (Table 4.1), we can estimate 600 quaternion sparse codes and then calculate 600 corresponding sparseness values. These values are illustrated by the histogram in Figure 4.10. One can see that the sparseness values of the SRCQW are averagely bigger than those of the SRWC and SRC (0.64 vs 0.62 and 0.57). More precisely, the largest sparseness value of SRCQW, SRWC, and SRC is 0.752, 0.716, and 0.665, respectively (Figure 4.10). Hence, we conclude that SRCQW provides a higher sparseness, which leads to better accuracy of classification.



(a) *SRC* **(b)** *SRWC* **(c)** *SRCQW*

Figure 4.10: *Sparseness histogram.*

In conclusion, the SRCQW, which is performed in the QW domain, promotes more sparsity of features than SRWC in the Wavelet domain. The proposed method provides the largest sparsity level, which contributes to the highest accuracy of classification.

4.6.8 Convergence rate

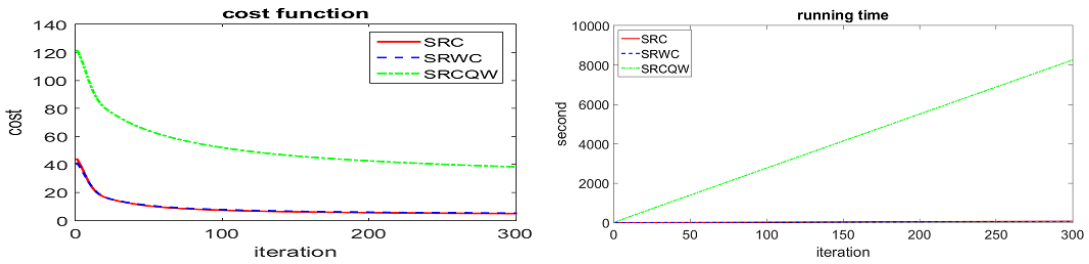


Figure 4.11: *Comparison of convergence rate and running time of the proposed SRCQW with SRC (Wright, A. Y. Yang, et al. 2008) and SRWC (Ngo et al. 2018) methods.*

To compare the convergence rate of the proposed SRCQW versus conventional SRC (Wright, A. Y. Yang, et al. 2008) and SRWC (Ngo et al. 2018) methods, we apply them on the AR face dataset. Figure 4.11 shows the cost functions and running time of the three methods on the interval of $[0, 300]$ iterations. One may observe that SRCQW has smaller rate of convergence compared with the other two. Moreover, the higher cost for SRCQW comes from the fact that it works in the 4D space of the AQ, and is a trade off for the very high accuracy of 99.85% and 99.6% (Table 4.2,4.3).

4.7 Discussion and Conclusion

The main contribution of this chapter is the development of the novel and efficient sparsity-promoting SRCQW method for image classification in the AQ, where features are described by the low-frequency QW sub-bands. Hence, the SRCQW method is the first SR-based one that uses information from the quaternion wavelet domain to solve the QWLasso minimization problem for classification in the 4D space of the AQ. Thus, we developed the novel gradient-based method QFISTA to calculate the sparse vector in the AQ defined over wavelets. Also, in QFISTA we develop an upper bound of $f(\mathbf{s})$ (Eq. 4.25), which simplifies the calculations.

The advantage of using quaternions is that they provide additional information, which comes from the direction of the vector of the quaternion in the 4D space of the AQ. This additional information increases the separability of the atoms, resulting in improvement of classification accuracy, which led to the achievement of 99.85% with YaleB dataset. To the best of our knowledge, no other method has reported such an accuracy of face identification. Another advantage of the SRCQW compared to other SRC in a transform domain such as DWT, is that the QWT has near shift-invariance and promotes higher sparsity of features than DWT, which make it particularly appropriate for data representation and discrimination. Hence, the QW domain makes the SR a very suitable representation learning method for SRC, providing robustness and high accuracy of classification.

Moreover, the SRCQW method is robust according to the number of training images as shown in Figure 4.8. For example, in the case of the COIL-100 dataset we use only 10 images per object for training and 62 for testing. This observation also shows the advantage of the SRCQW method over the NNs, which are very powerful and flexible tools for classification but need a large number of training samples to achieve high accuracy. In this sense we may state that, the NNs are useful classifiers if hundreds or at least tenth of thousands of training samples are available, while the SRCQW approach proved (see Table 4.3) to be optimal when medium-sized training dataset is available.

In the light of the FISTA algorithm (Beck et al. 2009) and due to the convexity of Eq. 4.24, the proposed QFISTA guarantees global convergence, while the methods in (Yi Xu et al. 2015; C. Zou et al. 2016) do not. The same statement holds for the NNs which usually need a number of parameters and functions to be correctly selected. On the other hand, the SRCQW method needs the selection of a single parameter (λ)

for which a few values should be tested (section 4.6.3).

A drawback of the novel SRCQW is its relatively high computational complexity, which is the trade off for the very high accuracy of 99.85%. Also, the method is inefficient for images where the background occupies a larger area than the target object.

In chapter 6, we will extend the application of the SRCQW method to skin lesion images. Further, this method promises a strong connection to the application of deep learning for the purpose of non-linear mapping.

CHAPTER 5

Contribution to the Convolutional Autoencoder Sparse Representation based Classification

Chapter content

5.1 Introduction

Over the last decades, sparse representation (SR) has been well studied and successfully applied on image classification (Wright, A. Y. Yang, et al. 2008) as well as many other image processing problems (Elad 2010). Recently, the trend of deep learning has also brought remarkable success to the research areas. Specifically, SR is known as one of the most successful approaches that produces a compact and simple representation of the data using a small number of meaningful features. Further, SR correlates with the visual neurons properties in the visual brain area (Olshausen and Field 1996).

However, the performance of SRC can be compromised when dealing with complex data whose samples from different classes may have high correlation. Another limitation of SRC is due to the way of using all the training images to form the dictionary, which is detrimental to the sparse code solver, especially in the case of big databases. For this reason, it is inevitable to develop an efficient approach to optimally represent the complex data by extracting the most meaningful features for classification purpose. Note that deep learning models have been successfully applied in many domains. More precisely, unsupervised autoencoders have been widely applied in several applications, especially in image processing (Papyan et al. 2017).

To further boost the classification accuracy for complex and big data, we propose, in this chapter, a novel CAE model regularized with sparsity constraint in the wavelet domain. We can consider this model as a hybrid approach between SRC (Wright, A. Y. Yang, et al. 2008) and CAE (Papyan et al. 2017) in the wavelet domain, which is employed to extract high discriminant features, being inspired from the proposed SRWC in chapter 3. The image features described by the complementary information of the low-frequency wavelet coefficients, which represents the most principal component of the image, are treated as the input to the model. Further, the loss function is regularized by a sparsity constraint in the latent space of the CAE. This allows the network to better learn the discrimination between the samples of different classes, and so, to result in a more accurate classification. Hence, this approach better classifies images with high variation in their contents, for example, the SVHN (Netzer et al. 2011) dataset, used in Section 5.4.

More precisely, the proposed method is inspired by the two latest methods, Deep Sparse Representation Classification (DSRC) (Abavisani et al. 2019) and Sparse Representation Wavelet based Classification (SRWC) (Ngo et al. 2018). However, unlike the

DSRC method (Abavisani et al. 2019), the proposed model contributes to the classification improvement by utilizing the wavelet features in order to exploit the sparsity of wavelets. Another key contribution is construction of a deep CAE architecture (see Table 5.1), which enhance the performance of classification and significantly decreases the number of network parameters. Moreover, the classification criterion from the residuals is based on a residual-based probabilistic rule (Wei et al. 2016) and not on minimum residuals criterion (Abavisani et al. 2019)

The remainder of this chapter is organized as follows. Section 5.2 presents related works. Then, the methodology of the proposed classification scheme is presented in Section 5.3. Subsequently, Section 5.4 discusses the experimental results before a conclusion in Section 5.5.

5.2 Related works

In the context of image classification, it is very important to find a suitable representation that captures the most meaningful properties of the data. Among the representation methods, SR has achieved great performances the signal and image processing literature (Elad 2010; Lu et al. 2014), where it is considered as a tool with the advantages of presenting high robustness to noise and other kinds of degradation (Wright, A. Y. Yang, et al. 2008).

A good representation of image features contributes as a main key to the success in any complex classification system. Among the explosive interests toward machine learning, deep learning models including VGG19 (Simonyan and Zisserman 2014), ResNet50 (He et al. 2016), Inception (Szegedy, Vanhoucke, et al. 2016), and WideResNet (Zagoruyko et al. 2016) have been developed and considered as powerful tools for learning data representation.

Taking advantage of sparsity, fusing neural networks with SR is a promising approach for image classification. In this respect, in (F. Li et al. 2018), the authors proposed a sparse AE-based model in order to improve the classification accuracy, by learning the SR using $\ell_{1/2}$ sparse regularization as a constraint on the hidden representation. Later, a CAE model fused with SRC (Wright, A. Y. Yang, et al. 2008), namely Deep Sparse Representation Classification (DSRC) model, was proposed. Its main idea is to a CAE to learn SR and estimate the sparse coefficients for classification. More precisely, embedding features are extracted from the input images by the encoder through a

non-linear mapping. These features are then fed to the sparse coding layer to find the sparse codes. Finally, the recovered embedding features are transferred into the decoder to reconstruct the images. In the training stage, the encoder-decoder and sparse coding are updated simultaneously. The resulting sparse codes are exploited to predict the class labels of test samples using the minimum reconstruction residual as mentioned in conventional SRC method (Wright, A. Y. Yang, et al. 2008). However, the conventional residual criterion is the drawback of DSRC method because it is not discriminant when the data between classes are highly correlated.

To further enhance the classification performance for large and complex datasets, a novel CAE model with sparsity regularization in the wavelet domain is introduced in the next section. The method benefits from the advantages of neural network, sparse representation and wavelet transform through the low frequency sub-bands.

5.3 Proposed method: Convolutional Autoencoder Sparse Representation Wavelet based Classification (CAESRWC)

In this section, we describe the newly proposed method, namely Convolutional Autoencoder based Sparse Representation Wavelet Classification (CAESRWC), to deal with classification problem of big and complex datasets. When comparing to recent deep learning-based methods (He et al. 2016; Simonyan and Zisserman 2014; Szegedy, Vanhoucke, et al. 2016; Zagoruyko et al. 2016), our method validates its enhanced performance as well as space efficiency with small number of network parameters.

The proposed method benefits from both the learning deep CAE, by providing the latent space with structured features in the wavelet domain, and the sparse coding, by providing sparse coefficients for the classification stage. Furthermore, the proposed method takes advantage of the wavelet domain, which promotes sparsity to enhance the classification reliability. As proved in (W. Zou and Yan Li 2007), using only the approximation coefficients from the low-pass sub-band, which causes better discrimination (see Section 5.3.1), will improve the classification accuracy.

The proposed classification scheme includes four main modules as illustrated in Fig. 5.1 and described below:

- The **Wavelet Transform** extracts the LL wavelet coefficients $\mathbf{X}$ of the original image $\mathbf{I}$ (including test or validation samples and training samples).

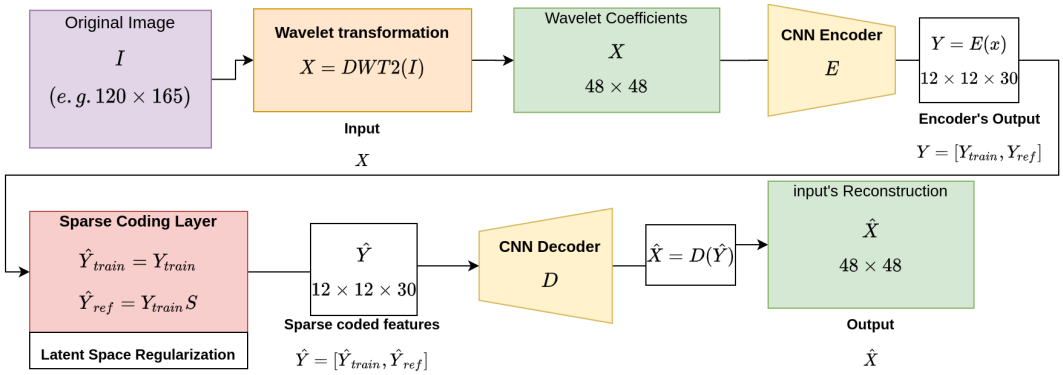


Figure 5.1: *End-to-end architecture of the proposed CAESRWC.*

- The **Encoder** extracts the encoding features $\mathbf{Y}$ in the latent space, with the extracted LL wavelet coefficients $\mathbf{X}$ as input.
- The **Sparse Coder** finds the sparse codes of the encoding features $\mathbf{Y}$ and yields the recovered encoding features $\hat{\mathbf{Y}}$ (concatenating the recovered encoding features and the training encoding features) to be fed into the decoder.
- The **Decoder** calculates from the sparse encoding $\hat{\mathbf{Y}}$ and a reconstruction $\hat{\mathbf{X}}$ close to the original input $\mathbf{X}$.

To train the autoencoder, the loss function is formed by minimizing both the reconstruction error and the sparse regularization terms. Then we exploit the estimated sparse codes of the encoding features to predict the class labels with the help of a residual-based probabilistic model.

Let $\mathbf{X}_{train} \in \mathbb{R}^{m \times n_{tr}}$, $\mathbf{X}_{val} \in \mathbb{R}^{m \times n_{val}}$, and $\mathbf{X}_{test} \in \mathbb{R}^{m \times n_{te}}$ be the given vectorized LL wavelet sub-bands of the training, validation, and testing data, respectively. Likewise, $\mathbf{Y}_{train} \in \mathbb{R}^{m_y \times n_{tr}}$, $\mathbf{Y}_{val} \in \mathbb{R}^{m_y \times n_{val}}$, and $\mathbf{Y}_{test} \in \mathbb{R}^{m_y \times n_{te}}$ are their corresponding encoding features. In our experiments (section 5.4.2), $\mathbf{X}_{train}$, $\mathbf{X}_{val}$, and $\mathbf{X}_{test}$ are three sub-packages extracted from the observed dataset for training, validation, and testing in the proportion of 0.8, 0.1, and 0.1, respectively ($n_{tr} = 0.8 * n, n_{val} = n_{te} = 0.1 * n$). The encoder input is defined as $\mathbf{X} = [\mathbf{X}_{train}, \mathbf{X}_{ref}] \in \mathbb{R}^{m \times (n_{tr} + n_{ref})}$ where "ref" refers to "*validation*" or "*test*".

5.3.1 Wavelet transform block

In this pre-precessing step, approximation wavelet sub-bands of the original images are extracted to feed the autoencoder. We convert each color image into gray-scale level

and apply the 5-level Haar wavelet transform. To further analyze the discrimination capability of each wavelet decomposition level, we conduct an analysis where the discrimination index of each level is computed using the images from two random classes of the observed data. In this analysis, we use the fisher ratio defined below as the discrimination index.

$$v = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{5.1}$$

where $\mu_1, \mu_2, \sigma_1, \sigma_2$ are mean and standard deviation values of the normalized approximation coefficients of images from the two classes.

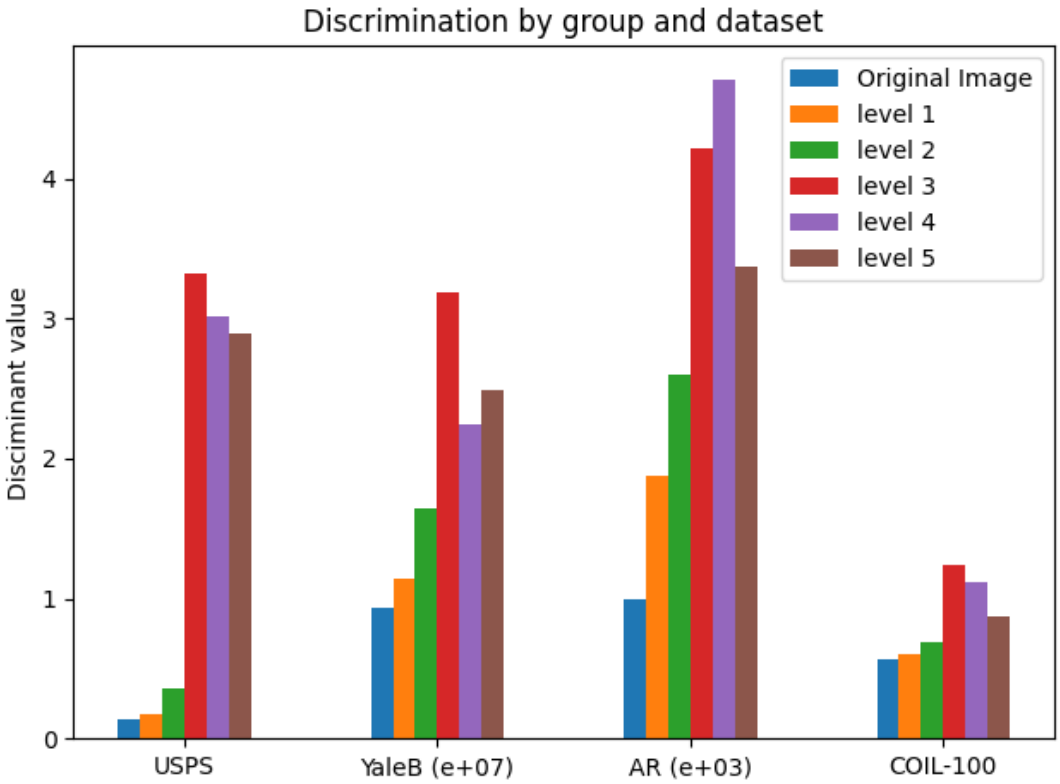


Figure 5.2: *Discriminant analysis based on different decomposition levels for various standard datasets.*

It can be observed from Fig. 5.2 that three resolution levels are sufficient to obtain a good discrimination for different standard image datasets.

5.3.2 Sparse coding block

The sparse coder plays an important role in this architecture. It estimates the sparse representation of the encoding features $\mathbf{Y} = [\mathbf{Y}_{train}, \mathbf{Y}_{ref}] \in \mathbb{R}^{m_y \times (n_{tr} + n_{ref})}$ in the latent

space by solving the following Lasso optimization problem:

$$\min_{\mathbf{S}} \|\mathbf{Y}_{ref} - \mathbf{Y}_{train}\mathbf{S}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 \ , \tag{5.2}$$

where $\mathbf{S} \in \mathbb{R}^{n_{tr} \times n_{ref}}$ is the sparse coefficients matrix and λ_1 is a positive regularization parameter that controls the sparsity penalty and the fidelity between the input $\mathbf{Y}$ and the estimated output $\hat{\mathbf{Y}}$ of the sparse coder. From (5.2), the estimated sparse encoding features $\hat{\mathbf{Y}}_{ref}$ can be considered as the output of a Fully Connected Network (FCN) with an input layer representing the encoded features vector $\mathbf{Y}_{train}$. Hence, $\hat{\mathbf{Y}}_{ref}$ can be computed as $\hat{\mathbf{Y}}_{ref} = \mathbf{Y}_{train}\mathbf{S} \in \mathbb{R}^{m_y \times n_{ref}}$. Forming the sparse coder's output as $\widehat{\mathbf{Y}} = [\widehat{\mathbf{Y}}_{train}, \widehat{\mathbf{Y}}_{ref}]$, problem (5.2) can be reformulated as:

$$\underset{\mathbf{S}}{min} \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 \ , \tag{5.3}$$

5.3.3 Loss function block

The proposed CAE-based sparse representation wavelet classification model will be trained based on an appropriate loss function which aims to:

- Learn a relevant representation of the input wavelet sub-bands $\mathbf{X}$ by a non linear reduction method, using the CAE instead of the linear PCA as performed in SRWC (Ngo et al. 2018). This results in reduced encoding wavelet features $\mathbf{Y}$ which allow to recover the reconstructed sub-bands $\widehat{\mathbf{X}}$ via $\hat{\mathbf{X}} = decoding(\hat{\mathbf{Y}}) = decoding(sparse\,coder(\mathbf{Y}))$. Thus, the loss $\mathcal{L}_{AE}$ due to the reconstruction error of the CAE between the input sub-band $\mathbf{X}$ and the reconstructed sub-band $\hat{\mathbf{X}}$ can be calculated as:

$$\mathcal{L}_{AE} = \left\|\mathbf{X} - \hat{\mathbf{X}}\right\|_F^2 \tag{5.4}$$

- Estimate the SR of the encoding features in the latent space by minimizing both the reconstruction error (i.e. first term in (5.3)) and the sparse regularization (i.e. second term in (5.3)). Thus, the loss related to the SR layer, $\mathcal{L}_{SR}$, is defined as follows:

$$\mathcal{L}_{SR} = \min_{\mathbf{S}} \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 \tag{5.5}$$

Therefore, the global loss function $\mathcal{L}_t$ of the proposed architecture composing of the CAE loss function $\mathcal{L}_{AE}$ and the SR one $\mathcal{L}_{SR}$ is defined as:

$$\mathcal{L}_t = \min_{\mathbf{S}} \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 \ + \lambda_2 \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F^2 \tag{5.6}$$

5.3.4 Classification stage

Once the training stage is completed, the estimated sparse codes are exploited to predict the labels of the test samples in the classification stage. More precisely, we utilize a residual-based probabilistic rule (Wei et al. 2016) that has been proven to be better compared to the conventional approach based on the truncated residual scheme (Wright, A. Y. Yang, et al. 2008) in the case of highly inter-correlated data. Firstly, we compute the residual r_c for each class c by:

$$r_c(\mathbf{x}_{test}) = \frac{\|\mathbf{y}_{test} - \mathbf{Y}_{train}\delta_c(\mathbf{s})\|_2^2}{\|\delta_c(\mathbf{s})\|_2^2}, \tag{5.7}$$

where $\mathbf{x}_{test}$ is the approximation sub-band of the observed sample in $\mathbf{X}_{test}$, while $\mathbf{y}_{test}$ is its embedding feature vector, and $\mathbf{s}$ is the corresponding sparse vector $\mathbf{s}$ in the sparse matrix $\mathbf{S}$.

By using the probability value associated with each residual r_c based on the softmax function, the label of the test sample $\mathbf{x}_{test}$ can be predicted by:

$$class\ (\mathbf{x}_{test}) = \arg\max_c(p_c) = \arg\max_c \left(\frac{e^{-r_c}}{\sum_{c=1}^{k} e^{-r_c}} \right), \tag{5.8}$$

where p_c denotes the probability that $\mathbf{x}_{test}$ belongs to class c, while k is the number of the classes.

From (5.8), it is evident to see that $0 \leq p_c \leq 1$ and $\sum_{c=1}^{k} p_c = 1$. However, the test sample $\mathbf{x}_{test}$ belongs to class c if the probability p_c is higher than a threshold (set to 0.99 in our experiments). Otherwise, its label is determined using the minimum residual in (5.7). Compared to the truncated residual criterion of the conventional SRC method (Wright, A. Y. Yang, et al. 2008), this probability judgement rule is a ratio of the basic residual term of SRC approach (Wright, A. Y. Yang, et al. 2008) to the ℓ_2-norm of the sparse code. This helps in overcoming the drawback of conventional SRC to classify highly inter-correlated datasets.

5.4 Experimental results

To evaluate the capability of the proposed method, we conduct various experiments and compare the results with recent state-of-the art image classification methods.

5.4.1 Experimental settings

	Layer	Output Shape	Kernel	Param#
Encoder	InputLayer	(None, 48, 48, 1)	∅	0
	Conv2D-1	(None, 48, 48, 10)	1×3×3×10	100
	Max-pool-1	(None, 24, 24, 10)	∅	0
	Conv2D-2	(None, 24, 24, 20)	1×3×3×20	1820
	Max-pool-2	(None, 12, 12, 20)	∅	0
	Conv2D-3	(None, 12, 12, 30)	1×1×1×30	630
	Max-pool-3	(None, 6, 6, 30)	∅	0
Sparse Layer	Dense-1	(None, 6, 6, 30)	∅	930
	Dense-2	(None, 6, 6, 30)	∅	930
Decoder	Conv2D-4	(None, 6, 6, 30)	1 1 1 30	930
	Upsampling1	(None, 12, 12, 30)	∅	0
	Conv2D-5	(None, 12, 12, 20)	3×3×3×20	5420
	Upsampling2	(None, 24, 24, 20)	∅	0
	Conv2D-6	(None, 24, 24, 10)	3×3×3×10	1810
	Upsampling3	(None, 48, 48, 10)	∅	0
	Conv2D-7	(None, 48, 48, 1)	3×3×3×1	91
Param	**Total parameters: 12, 661**			
	Trainable parameters: 12, 661			

Table 5.1: *Description of the proposed model's parameters*

We carry out the training process of the proposed model using Tensorflow 2.0, and NVIDIA Tesla T4 GPU. Our model is trained using the momentum Adam optimizer with the learning rate $1e-3$ while applying a decay of 0.9. To train this model, a two-stage approach is considered. In the first one, considered as a pre-trained stage, the training process is launched without using the sparse coding layer, like a traditional CAE model, in 100 epochs. Then, in the second stage, the overall model including the sparse coding layer is trained in 900 epochs. The summary of the end-to-end model can be found in Table. 5.1. The number of neurons of the input layer corresponds to the dimension of the input approximation wavelet sub-band. We use the kernel size of 1×1 for the third convolution layer while the size of 3×3 is used for the rest convolution layer. To deal with overfitting, we use the dropout and random permutation cross-validation in our experiments. Finally, the parameters λ_1 and λ_2 in (5.6) are set to 10 and 1, respectively.

5.4.2 Datasets

In this section, we evaluate our method against state-of-the art SR-based methods (SRWC (Ngo et al. 2018), FDDL (M. Yang, L. Zhang, Feng, et al. 2011), LC-KSVD2 (Jiang et al. 2013)), and DSRC(Abavisani et al. 2019)). Two digits datasets (USPS (Hull 1994) and SVHN (Netzer et al. 2011)), three face datasets (AR face (Martinez 1998), YaleB (Georghiades et al. 2001) and UMDAA-01 (Heng Zhang et al. 2015)), one object dataset COIL-100 (Nene et al. 1996), and AR gender dataset (Martinez 1998) are considered in our experiments. Some examples of each dataset are shown in Fig. 5.3. These datasets can be briefly described as follows:

(a) *USPS (Hull 1994)* **(b)** *SVHN (Netzer et al. 2011)* **(c)** *UMDAA-01 (Heng Zhang et al. 2015)*

(d) *YaleB (Georghiades et al. 2001)* **(e)** *AR (Martinez 1998)* **(f)** *Coil-100 (Nene et al. 1996)*

Figure 5.3: *Some data samples from the six employed datasets.*

- **USPS (Hull 1994)** is the handwritten digits dataset, which consists of 7291 training gray-scale images of ten digits (0-9).A subset of 2000 samples is considered in our experiment. Hence, 1600, 200, and 200 samples are randomly selected for the training, validation, and testing, respectively.

- **Street view house numbers (SVHN) (Netzer et al. 2011)** is a real-world dataset used for object recognition methods. There are over 600,000 labeled real-world images of house numbers obtained from Google Street View images in this dataset. A subset of 2000 samples corresponding to 10 classes has been rcpnsider. Hence, 160, 20, and 20 samples per class are used for the training, validation, and testing, respectively.

- **AR face (Martinez 1998)** contains over 4,000 color images relating to 126 people's faces (70 men and 56 women). In our experiments, a subset of 2000

samples corresponding to 100 classes are considered, where 1600, 200, and 200 samples are used for the training, validation, and testing, respectively.

- **YaleB face (Georghiades et al. 2001)** consists of 5760 single light source pictures of 10 subjects observed under 576 viewing conditions. In our experiments, we retained a subset of 2000 samples where 1600, 200, and 200 samples are randomly selected for the training, validation, and testing, respectively.

- **UMD mobile face (Heng Zhang et al. 2015)** consists of 750 front-facing camera videos of 50 users captured by smartphones. In our experiments, we randomly selected 2500 images corresponding to 50 classes where 2000, 250 and 250 samples are used for training, validation, and testing, respectively.

- **COIL-100 (Nene et al. 1996)** consists of 7,200 color images corresponding to 100 objects. Then, 4800, 1200 and 1200 samples are randomly selected for the training, validation, and testing, respectively.

- **AR gender (Martinez 1998)** is the last considered dataset which consists of 2,600 face images corresponding to 50 males and 50 females taken under 26 viewing conditions. In our experiments, 2,000, 300 and 300 images are randomly selected for training, validation, and testing, respectively.

5.4.3 Performance and comparison

The overall accuracy results of the proposed methods and contemporary ones are reported in Table. 5.2 where the best values are highlighted in bold.

	Method					
Acc. (%)	**SRC**	**FDDL**	**SRWC**	**LC-KSVD2**	**DSRC**	**CAE SRWC**
USPS	87.78	91.34	95.45	87.45	96.25	**96.82**
SVHN	15.71	22.54	28.21	35.31	67.75	**68.24**
ARface	97.61	96.16	**98.39**	97.70	98.12	98.37
ARgender	93.0	94.0	94.2	86.8	96.48	**96.54**
YaleB	97.54	97.52	98.06	97.80	97.20	**98.35**
UMDAA-01	79.00	81.22	85.29	84.82	93.39	**95.10**
COIL 100	91.16	88.22	92.29	91.42	91.12	**92.35**

Table 5.2: *Classification accuracy (%).*

One can see from Table. 5.2 that the proposed method CAESRWC outperforms the state-of-the art methods for most of the employed datasets. The proposed method

ranks second with AR face dataset (Martinez 1998) and is very close to the SRWC method (Ngo et al. 2018) (i.e. 0.02%).

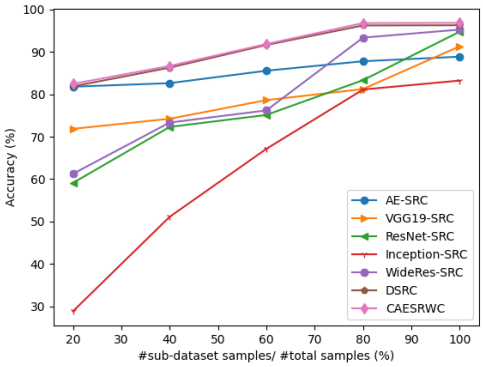
To further validate the advantage of the proposed method over deep neural networks, which are well known to be effective only with large scale training data, we propose to analyze the accuracy of different deep learning based methods as a function of the number of training samples. For that, four subsets are formed by randomly choosing 20%, 40%, 60% and 80% of the sample of each dataset. Then, we separate each subset into training/validation/test samples and conduct the experiments on different classification methods. Fig. 5.4 illustrates the accuracy of these methods as a function of the training size on six different datasets.

One can see from Fig. 5.4 that the modern deep learning based models are sensitive to the training size. They perform a significant drop in accuracy when the training size is relative small. This is an inevitable drawback of moden trained models. In the meantime, the proposed method is superior to the others including DSRC method and appears to be more robust to the number of training samples.

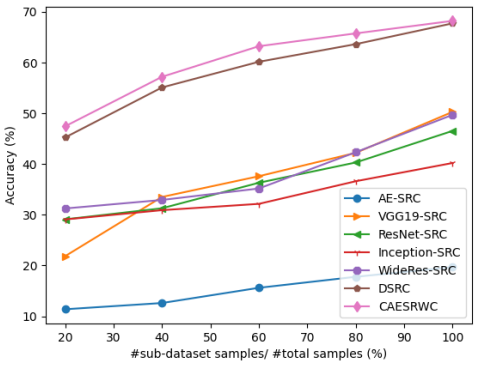
Finally, a comparison between different deep learning based methods in terms of the network parameters has been studied as shown in Table 5.3. The number of parameters of the proposed method are much smaller than the other methods. Obviously, the reduced number of parameters explains the good behavior of the proposed method with limited amount of training samples. Moreover, a small sized model allows to achieve gain in terms of storage memory.

	VGG	**ResNet**	**Inception**	**Wide ResNet**	**DSRC**	**CAE SRWC**
#param	138M	25.6M	23.8M	8.8M	24.5K	**12.7K**

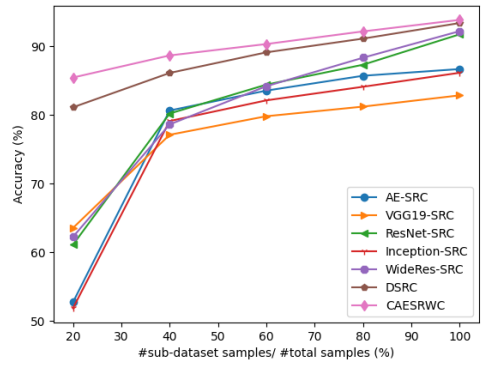
Table 5.3: *Comparison of the number of network parameters between different deep learning methods.*



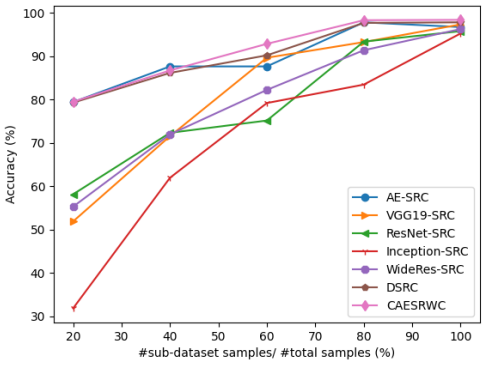
(a) *USPS (Hull 1994)*



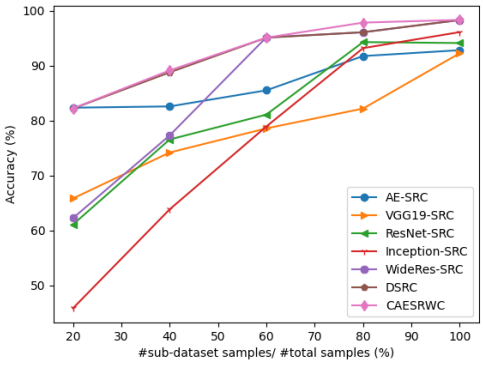
(b) *SVHN (Netzer et al. 2011)*



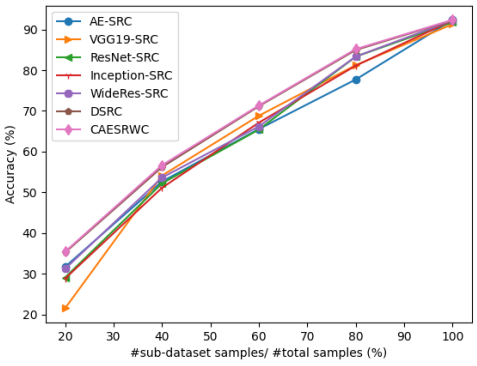
(c) *UMDAA-01 (Heng Zhang et al. 2015)*



(d) *YaleB (Georghiades et al. 2001)*



(e) *AR (Martinez 1998)*



(f) *Coil-100 (Nene et al. 1996)*

Figure 5.4: *Effects of the size of the employed subsets of data on the accuracy for the different deep learning methods.*

5.5 Conclusion

In this chapter, a novel classification method fusing the sparse representation and deep learning model has been proposed in the wavelet domain. More precisely, the method relies on a convolutional autoencoder architecture and a sparse latent layer applied to the low-pass sub-bands. Extensive number of experiments are carried out on different datasets and reveal the advantages of the proposed method over other recent state-of-the-art deep learning based approaches. For the future work, instead of using only the approximation sub-band, we will investigate the effect of high-pass sub-bands as well as other wavelet transforms rather than Haar wavelet on the classification.

CHAPTER 6

Application to Skin Lesion Diagnosis

Chapter content

6.1 Introduction

Melanoma is a type of skin lesion that mostly develops in the pigment-producing cells, also called melanocytes which are responsible for the production of melanin. By leading to 9000 deaths a year and accounting for about 75% of deaths associated with skin cancer, it is considered as the most dangerous form of skin cancer(Jerant et al. 2000). The American Cancer Society estimated that about 105540 new cases of melanomas would be evaluated in the United States in 2019 (Siegel et al. 2019). Fortunately, if melanoma is diagnosed early, it can be cured properly to enhance the survival rate of patients (Balch et al. 2001).

In order to improve melanoma detection the dermoscopy technique was developed. Dermoscopy is a noninvasive skin imaging technique of acquiring a magnified and illuminated visualization of a subsurface structure of skin to increase the clarity of its spots (Binder et al. 1995). By eliminating surface reflection of skin and using cross-polarized light, the dermoscopy enhances the visual effect of deeper levels of skin and hence provides more details to increase the discrimination between lesions. Dermoscopy assessment is widely used in the clinical diagnosis of melanoma and obtains much higher accuracy rates than evaluation by naked eyes (Silveira et al. 2009). On the other hand, early detection of melanoma is classically evaluated with physical examinations of skins based on morphological features such as atypical pigment network, dots, globules, streaks, pseudopods, blue-whitish veil, regression pattern, atypical vascular pattern, structureless areas. For example, the presence of dots and globules is associated with increased probability of melanoma. This holds especially if more than three of them are black or dark brown colors, and are located on the periphery (Sirakov, Mete, et al. 2015). The problem is that sometimes dots are difficult to detect even by experienced dermatologist equipped with a dermoscope (Sirakov, Mete, et al. 2015). Fig. 6.1 shows some examples of skin lesion dots in dermoscopy images. Moreover, the manual inspection from dermoscopy images is still complex, time-consuming, fault-prone and subjective due to the challenges with various image characteristics (variation in lesion size, color or shape, presence of artifacts, etc) (Binder et al. 1995; Vestergaard et al. 2008). Hence, there is high demand for the development of reliable automated classification approaches in order to remedy these limitations.

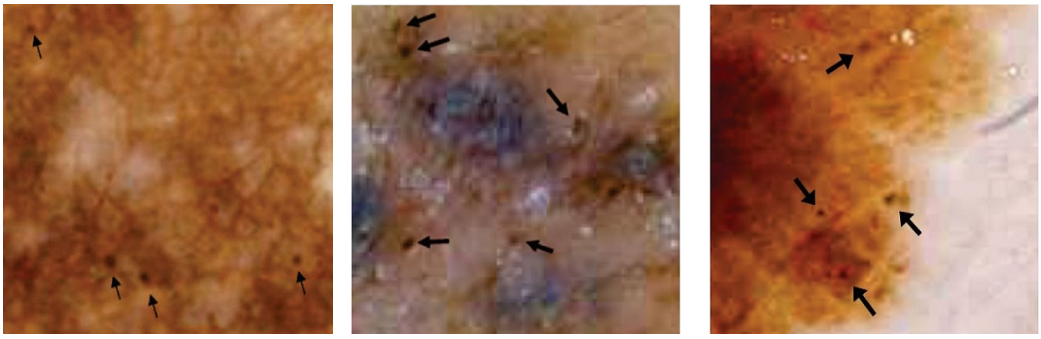
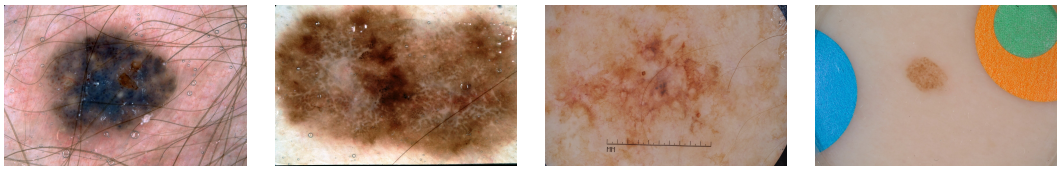


Figure 6.1: *Example of skin lesion dots (with ground truth (Sirakov, Mete, et al. 2015) indicated by arrows) in dermoscopy images.*

Automated skin lesion classification from dermoscopy images is a challenging task due to many issues. First, the large intra-class variation of melanomas in terms of color,

texture, shape, size and location in the dermoscopy images as well as the high degree of visual similarity between melanoma and benign lesions make it difficult to discriminate melanomas from benign skin lesions. Second, the relatively low contrasts and obscure boundaries between skin lesions (especially at their early stages) and normal skin regions make the automated classification task even harder. Finally, the presence of artifacts such as hairs, veins, air bubbles, or color reflections may blur or occlude the skin lesions (Fig. 6.2).



(a) *Skin hairs* **(b)** *Air bubbles* **(c)** *Ruler marks* **(d)** *Color charts*

Figure 6.2: *Example of skin lesions from the ISIC2017 dataset with various artifacts.*

A lot of efforts have been dedicated to deal with this problem. Early investigations apply low-level hand-crafted features to distinguish melanomas from benign skin lesions, including shape (Mete et al. 2012), color (Sirakov, Ou, et al. 2015), and texture (Ballerini et al. 2013). Later, feature selection algorithms have been proposed to define proper features and employ them to improve the recognition performance (Ganster et al. 2001; Sirakov, Mete, et al. 2015). However, these hand-crafted features are incapable of dealing with the large intra-class variation of melanoma and the visual similarity between melanoma and benign lesions. On the other hand, in designing methods for a computer-aided diagnosis (CAD) system which is generally composed of sequential processes including pre-processing, segmentation, and classification, some researchers proposed to firstly perform segmentation and then to recognize the melanomas (Ganster et al. 2001; Mete et al. 2012; Sirakov, Mete, et al. 2015; Sirakov, Ou, et al. 2015). However, this approach deals with two main challenges: (i) each process is complex with the need to tune a set of parameters, specific to a given dataset; (ii) the performance of each process depends on the previous one, and the errors are built up throughout the system (Rastgoo, Lemaitre, et al. 2016). In the recent years, the deep neural networks (NNs) have gained the popularity in the field of automated melanoma classification but they always need a large amount of training samples, which is not always available in real-world tasks.

In the present chapter, the novel SRCQW method, which is proposed and presented

in Chapter 4, is applied on skin lesion images classification. Further, we also studied the SRCQW using the high-frequency quaternion sub-bands along with the low-frequency one and the mixed version between them. This study allows to analyse and determine the sub-band appropriate for the classification of skin lesion images.

In every domain, we conducted experiments to classify the skin lesion images from the ISIC2017 (N. C. Codella et al. 2018) and ISIC2019 (N. C. Codella et al. 2018; Combalia et al. 2019; Tschandl et al. 2018) datasets. The extensive number of experimental results demonstrated that the proposed method is competitive with several state-of-the-art methods including deep learning.

The rest of the chapter is organized as follows: Section 6.2 reviews related works; Section 6.3 validates the SRCQW on the skin lesion datasets ISIC2017 (N. C. Codella et al. 2018) and ISIC2019 and compares the obtained results with several contemporary methods; Section 6.4 concludes the chapter by listing the contributions, with a discussion on the advantages and bottlenecks.

6.2 Related works

In the past decade, numerous approaches have been proposed for automatic diagnosis of skin lesions. A summary of these methods and their properties can be found in the article of (Korotkov et al. 2012). Unfortunately, a fair comparison among the state-of-the-art presented methods is not possible due to lack of a benchmark and common datasets (Korotkov et al. 2012; Rastgoo, Garcia, et al. 2015). Jaworek-Korjakowska (Jaworek-Korjakowska 2012) proposed a classification system based on the most clinically used ABCD rule for melanoma detection. More specifically, features (i.e. asymmetry, border irregularity, amount of colors and diameter) of every lesion are computed and utilized to determine whether it is a melanoma or benign lesion. Other features in use are shapes, colors, dots and texture. Some of the commonly used classifiers are Support Vector Machine (SVM) (Abuzaghleh et al. 2014; Mete et al. 2012; Sirakov, Mete, et al. 2015; Sirakov, Ou, et al. 2015), k-nearest neighbors (KNN) (Ballerini et al. 2013; Ganster et al. 2001), AdaBoost (Ruela, Barata, and Marques 2013; Ruela, Barata, Mendonça, et al. 2013), Local Binary Pattern (LBP) (Riaz et al. 2014), Bag-of-feature (BoF) (Barata et al. 2013), and ensemble approach (Rastgoo, Garcia, et al. 2015). In particular, (Abuzaghleh et al. 2014) proposed the use of SVM based on 2D Fast Fourier Transform (FFT2) and Discrete Cosine Transform (DCT). (Ruela, Barata, and Marques

2013; Ruela, Barata, Mendonça, et al. 2013) applied an AdaBoost classifier to compare the role of shapes and colors for classification. (Barata et al. 2013) proposed the use of a BoF model including colors and gradient features. (Riaz et al. 2014) introduced a variation of LBP descriptor and combined it with color features for the dermoscopy image classification. (Rastgoo, Garcia, et al. 2015) compared the effects of various colors, shape and texture features using ensemble approaches. The features were extracted from previously segmented area and data space over-sampling (DOS) was used instead of random over-sampling (ROS). Noroozi and Zakerolhosseini (Noroozi et al. 2016), for the first time, developed a novel method for detecting basal cell carcinoma tumor using Z-transform features as a combination of two or three Fourier transform features.

Later, with the success of sparse coding approach, a number of methods have been developed. (Rastgoo, Lemaitre, et al. 2016) proposed a melanoma classification framework based on sparse coding without the pre-processing or lesion segmentation step. More precisely, Random Forests classifier and SR were utilized with the help of the features including SIFT, Hue and Opponent angle histograms, and RGB intensities. (N. Codella et al. 2015) combined deep learning, sparse coding and SVM algorithms to better characterize the lesions for melanoma classification. (Yao et al. 2016) proposed a multi-view joint SR framework for melanoma detection. In this method, the local texture and color features were extracted and then the SR of multiple features was jointly learned with a discriminative dictionary. Recently, Moradi et al (Moradi et al. 2019) proposed a framework for melanoma segmentation and classification based on kernel SR. For this purpose, selected features are represented in a high dimensional feature space by a kernel-based learned dictionary and discriminative sparse codes.

Inspired from the recent growth of the neural network (NN) and deep learning applications for solving scientific, medical and industrial problems, a number of NNs were launched in the literature (Astudillo et al. 2020; Yuexiang Li et al. 2018; Sousa et al. 2017; J. Zhang et al. 2019). In (Astudillo et al. 2020) the authors developed a convolutional NN (CNN) and applied it with noisy Stochastic Gradient Descent (SGD) and Adam learning methods on the ISIC2018 skin lesion dataset. Its subset ISIC2017 (N. C. Codella et al. 2018) was classified by the NNs presented in (Yuexiang Li et al. 2018; Sousa et al. 2017; J. Zhang et al. 2019).

The next section presents the application of the SRCQW proposed in chapter 4 to skin lesion images classification, which is the main purpose of this chapter.

6.3 Application of the SRCQW to skin lesion images classification

In order to analyse and determine the sub-band appropriate for the classification of skin lesion images, the proposed SRCQW method is coded in Matlab and applied to each of the quaternion sub-bands $\left\{\dot{LL}, \dot{LH}, \dot{HL}, \dot{HH}\right\}$ and every pair from the set to classify skin lesion images to melanoma and benign lesion in the large public datasets, namely ISIC2017 (N. C. Codella et al. 2018) and ISIC2019 (N. C. Codella et al. 2018; Combalia et al. 2019; Tschandl et al. 2018). Then we compare its results with several contemporary methods in the field (Yuexiang Li et al. 2018; Rebouças Filho et al. 2018; Sousa et al. 2017; J. Zhang et al. 2019) as well as with the SRWC (proposed in chapter 3) and the CAE-SRWC (proposed in chapter 5). Note that, for CAE-SRWC method (Nguyen et al. 2020), we used the same experimental settings as mentioned in section 5.4.1, but using Pytorch 1.7.1 instead of Tensorflow 2.0 and NVIDIA RTX2060 super GPU instead of NVIDIA Tesla T4 GPU. While (Rebouças Filho et al. 2018) proposed to automatically classify melanoma from dermoscopy images using structural co-occurrence matrix of main extracted frequencies, (Sousa et al. 2017), (Yuexiang Li et al. 2018), and (J. Zhang et al. 2019) presented different convolutional neural networks to address the classification problem. In adition, we compare the highest results from the ISIC2017 challenge[1] with those obtained by the novel SRCQW.

Note, (C. Zou et al. 2016) recommends using λ values from the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Experimenting with these values, we found that the novel SRCQW method obtains its higher performance of skin lesion binary classification with $\lambda = 10^{-3}$, which we use in our further experiments.

In our experimental setup, we use Monte Carlo cross-validation (Dubitzky et al. 2007) as described in section 4.6.1, to validate the capability of the proposed method. We randomly select the training and test sets from the dataset. The results are averaged over the k splits. The advantage of this set-up is that the variance of the split sample error estimate can substantially be reduced. Also, the proportion of the training-test random splits does not depend on the number k (Molinaro et al. 2005). In this chapter, k is set to 10.

[1] https://challenge.isic-archive.com/landing/2017

6.3.1 Dataset

We conducted extensive experiments to evaluate the novel SRCQW method on one of the largest public datasets of skin lesion images, ISIC2017 (N. C. Codella et al. 2018). The dataset consists of 2229 benign and 521 malignant melanoma images equipped with the gold standard diagnosis. The ground truth is held out by the ISBI 2017 (N. C. Codella et al. 2018) organizer for independent evaluation. Further a challenge competition was organized (N. C. Codella et al. 2018). Multiple teams participated providing classification results by contemporary NNs.

The numbers of melanoma and benign lesions in this dataset are quite imbalanced. In this regard, it will bias the dictionary generation of our method. In order to remedy the problem, we randomly select 1708 malignant images from the ISIC Archive[2] resource to obtain a new dataset that we referred to as 'ISIC2017+' dataset. This latter contains 2229 malignant and 2229 benign skin lesion images. Note that the authors in (J. Zhang et al. 2019) (Table 6.3), also collected additional dermoscopy images (1320 including 466 melanoma, 822 nevus images, and 32 seborrheic keratosis images) from the ISIC Archive to enlarge the training dataset. Likewise, in (Yuexiang Li et al. 2018) (Table 6.3), the authors increased the dataset by using data augmentation to obtain 7480 melanoma, 10976 nevus images, and 5080 seborrheic keratosis images. To validate the capability of efficiently classifying middle size dataset, we randomly select 1115 malignant and 1115 benign skin lesion images from the ISIC2017+ dataset to conduct the experiments. Every image is supplied with a ground truth diagnosis to benign or malignant lesion. Examples of the dataset are shown in Fig. 6.3. We also report in Fig. 6.4 an example of a skin lesion image decomposed by the QWT to 16 wavelet sub-bands.

We observed that most of the images selected for the set of experiments possess a large background around the lesion. In order to balance the images, we cropped them, around the center, 80% of every image and used the cropped area for classification. We briefly report the description of the used dataset in Table 6.1 along with the number of skin lesion image features used from a single image to create a dictionary atom.

Further, the proposed SRCQW was validated on the ISIC2019 dataset, which is an extension of the ISIC2017. There are 25331 dermoscopic images of eight different skin lesion classes, which are originally taken from the the HAM10000 dataset (Tschandl et al. 2018), the BCN20000 dataset (Combalia et al. 2019), and the MSK dataset (N. C.

[2] https://www.isic-archive.com/

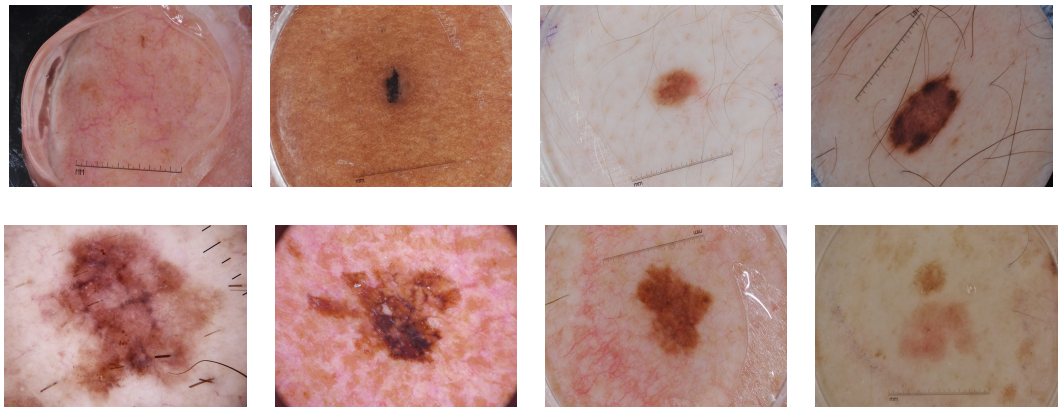
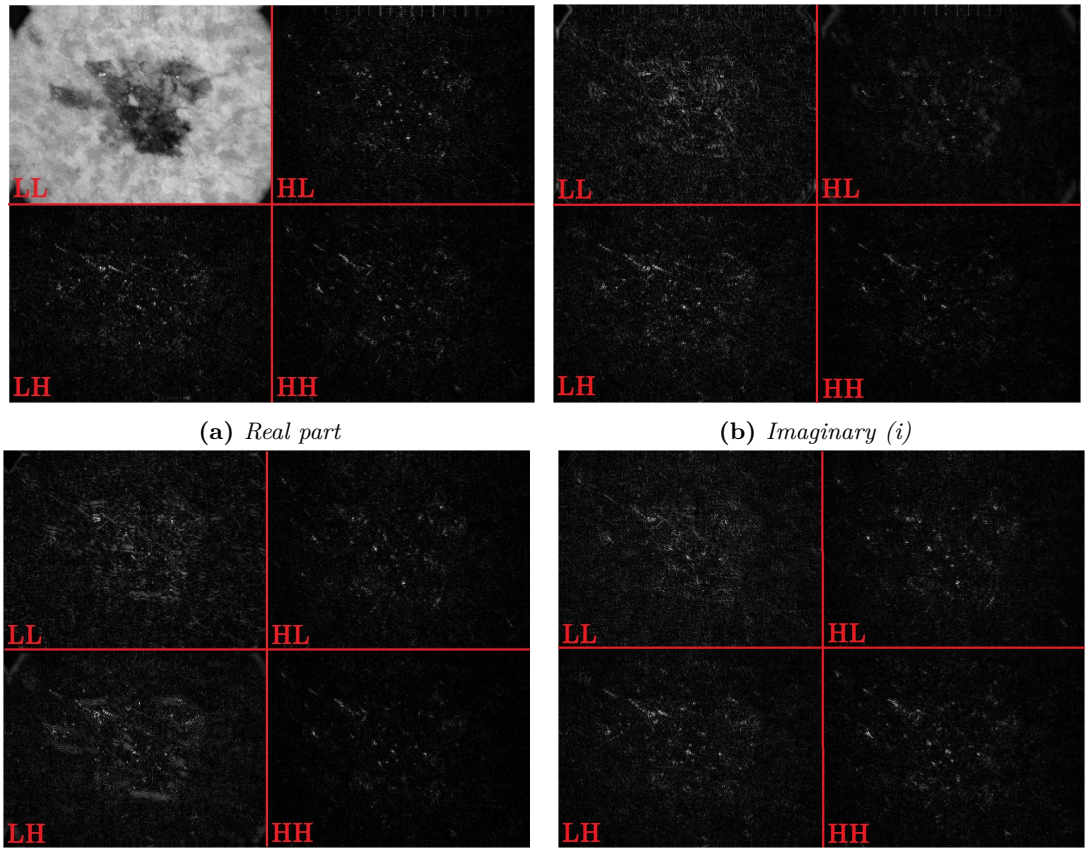


Figure 6.3: *Examples of some skin lesion images from ISIC2017+. The first and second rows are benign and melanoma lesions, respectively.*



(a) *Real part* **(b)** *Imaginary (i)*

(c) *Imaginary (j)* **(d)** *Imaginary (k)*

Figure 6.4: *Example of a skin lesion image decomposed by the QWT to 16 wavelet sub-bands. The gray level version of the original image is shown at the upper left corner of the real part.*

Table 6.1: *Description of the ISIC2017+ dataset used in the case of single frequency and mix frequencies. Columns 3 and 4 show the number of training samples and number of test samples, respectively.*

	Cropped image size	**#Training**	**#Test**	**Feature dim**
$\dot{LL}/\dot{LH}/\dot{HL}/\dot{HH}$	256×192	$N = 2007$	223	$M_1 = 3600$
$\dot{LL}\&\dot{LH}/\dot{LL}\&\dot{HL}/\dot{LL}\&\dot{HH}$	256×192	$N = 2007$	223	$M_2 = 4000$

Codella et al. 2018). We randomly selected 1000 images from ISIC2019, having 500 melanomas and 500 benign. In the pre-processing stage, we applied the Laplacian filter ($\Delta I + 2I$, where $I(x, y)$ is the image function) on the original images. Next, we cropped them as done with the above ISIC2017+.

6.3.2 Evaluation metrics

For the purpose of comparison, we apply seven metrics to evaluate the classification effectiveness of the proposed SRCQW method, including accuracy (AC), specificity (SP), sensitivity (SE), miss rate (MR), precision (PR), F1 score (F1), and Youden's J statistic (J). The metrics are defined as:

$$\begin{aligned} AC &= \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{fn} + N_{tn}}, \\ SE = \frac{N_{tp}}{N_{tp} + N_{fn}}, \quad SP &= \frac{N_{tn}}{N_{tn} + N_{fp}}, \\ PR = \frac{N_{tp}}{N_{tp} + N_{fp}}, \quad F1 &= 2\frac{PR \cdot SE}{PR + SE}, \\ MR = 1 - SE, \quad J &= SE + SP - 1, \end{aligned} \tag{6.1}$$

where N_{tp}, N_{fp}, N_{fn} and N_{tn} denote the number of true positive, false positive, false negative and true negative, respectively. A melanoma image is considered as a true positive if its SRCQW evaluation is melanoma; otherwise it is regarded as a false negative. A benign image is considered as a true negative if its diagnosis is benign; otherwise it is regarded as a false positive.

6.3.3 Results and discussion

In our experiments, we randomly select 90% of the images for training, while the rest is used for testing (Table 6.1). We execute the code ten times in Matlab environment using the same number of training samples for all runs, but randomly select different training set for every run. Under the above settings we experimented the SRCQW method with every quaternion sub-band $\left\{\dot{LL}, \dot{LH}, \dot{HL}, \dot{HH}\right\}$ and every pair of them.

The average classification rates from every set of these experiments are reported in Table 6.2. Note that the results with pairs of quaternion sub-bands that do not contain $\dot{LL}$ are not shown due to the SRCQW poor performance.

Table 6.2: *Application of the SRCQW method to skin lesion images with performance comparison of different quaternion sub-bands on ISIC2017+ dataset. Best results for each index are marked in bold.*

	AC (%)	SP (%)	SE (%)	PR (%)	F1 (%)	J (%)
$\dot{LL}$	**88.19**	**98.09**	78.28	**97.63**	**86.86**	**76.37**
$\dot{LH}$	52.25	58.56	45.95	52.59	49.04	4.51
$\dot{HL}$	55.86	57.66	54.05	56.07	55.05	11.71
$\dot{HH}$	62.16	57.66	66.67	61.16	63.79	24.33
$\dot{LL}\&\dot{LH}$	66.22	50.45	**79.58**	62.33	70.82	30.03
$\dot{LL}\&\dot{HL}$	62.61	52.25	72.97	60.45	66.12	25.22
$\dot{LL}\&\dot{HH}$	60.36	46.85	73.87	58.16	65.08	20.72

Table 6.2 shows the skin lesion classification results of the novel SRCQW method using one or two quaternion sub-bands for classification. It is evident that SRCQW with the low-frequency sub-band $\dot{LL}$ provides the best results with huge difference compared to the others in almost metrics except in sensitivity, where classification with $\left\{\dot{LL}\&\dot{LH}\right\}$ gives the highest value of sensitivity. In other words, high values of sensitivity in Table 6.2 indicate that mixed frequencies with low-frequency sub-band $\dot{LL}$ provide the highest rate of melanoma recognition with 'ISIC2017+' dataset. Hereafter, for comparison purposes we report results obtained by SRCQW with the low-frequency quaternion sub-band $\dot{LL}$.

Table 6.3: *Comparison of the proposed SRCQW with contemporary methods. Best results for each index are marked in bold.*

	AC (%)	SP (%)	SE (%)	PR (%)	F1 (%)	J (%)
(Sousa et al. 2017)	84.7	90	63.3	-	-	53.3
(Yuexiang Li et al. 2018)	85.2	93.3	50.4	-	-	43.7
(Rebouças Filho et al. 2018)	**89.93**	92.15	**89.93**	91.29	**90**	**82.08**
(J. Zhang et al. 2019)	85	89.6	65.8	-	-	55.4
SRWC	63.68	45.29	82.06	60	69.32	53.04
CAE-SRWC	76.68	78.5	75	79.09	76.99	53.5
SRCQW	88.19	**98.09**	78.28	**97.63**	86.86	76.37

To validate the classification capabilities of the novel SRCQW method, we compare it with six contemporary methods, including the two proposed methods SRWC (Chapter 3) and CAE-SRWC (Chapter 5), on the 'ISIC2017+' dataset described in section 6.3.1. One may observe in Table 6.3 that the newly proposed SRCQW method significantly outperforms the competitors in specificity (98.09%, making a substantial improvement

up to 4.79%) and precision (97.63%, making a significant improvement up to 6.34%). Its accuracy is relatively high (88.19$) and is quite close to the highest one. These results prove that the proposed SRCQW is the most accurate in recognizing benign lesions if the low-frequency quaternion sub-band $\dot{LL}$ is used. This is a useful advise to dermatologists, because it leads to decrease of wrongly treated benign lesions as melanoma, which decreases the total cost of treatment. Although Youden's index (J) of the proposed SRCQW ranks second behind the one of (Rebouças Filho et al. 2018), its value is relatively large (76.37%), indicating that the proposed method is well balanced in diagnosing both benign and melanoma images. The SRWC (Ngo et al. 2018) does not generally perform well except for sensitivity (82.06%), which means that the method is very sensitive to the melanoma detection. It can be observed that the CAE-SRWC (Nguyen et al. 2020) obtains even results in every metrics, which shows its balance in diagnosing both melanoma and benign images.

As mentioned above, each experiment is executed repeatedly ten times to obtain the average, maximum, and minimum results. The maximum and minimum results obtained by the SRCQW method with $\dot{LL}$ in each of the five metrics are reported in Table 6.4. These results are verified by the corresponding confusion matrix presented in Table 6.5, which shows the number of correct and wrong classified images per class. One can see in Table 6.5 that the SRCQW correctly classify 92 images as melanoma, but misclassify 19 melanoma images as benign. Similarly, it accurately labels 110 images as benign but wrongly labels 1 benign image as melanoma.

Table 6.4: *Maximum and minimum (%) results of the proposed SRCQW method on the ISIC2017+ dataset.*

	AC (%)	SP (%)	SE (%)	PR (%)	F1 (%)
Max	90.54	99.1	82.88	98.89	89.76
Min	85.59	96.4	72.97	95.65	83.51

Table 6.5: *An example confusion matrix obtained by the SRCQW method on the ISIC2017+ dataset.*

	Condition positive	Condition negative
Predicted condition positive	$N_{tp} = 92$	$N_{fp} = 1$
Predicted condition negative	$N_{fn} = 19$	$N_{tn} = 110$

As mentioned above, we conducted experiments on classifying the ISIC2017+ skin lesion dataset with the newly proposed SRCQW method. To validate its classification capabilities, the SRCQW is compared (Table 6.6) with CAE-SRWC (Nguyen et al. 2020) and the results of the top 10 deep NN-based classifiers (according to the leaderboard

Table 6.6: *Classification comparison with the top 10 results of the ISIC2017 skin lesion classification challenge (N. C. Codella et al. 2018). The two best results for each metric are marked in bold.*

	AC (%)	SP (%)	SE (%)	MR (%)	J (%)
RECOD Titans	**87.2**	95	54.7	45.3	49.7
USYD-BMIT1	85.8	96.3	42.7	57.3	39
CSUJT	82.8	85.1	73.5	26.5	**58.6**
MPG-UCIIIM	82.3	**99.8**	10.3	89.7	10.1
UoG-MLRG	84.5	96.5	35	65	31.5
IHPC-NSC	83	92.5	43.6	56.4	36.1
UFdMG	82.7	90.1	52.1	47.9	42.2
CVI	84.3	95.7	37.6	62.4	33.3
icuff1	83	**99**	17.1	82.9	16.1
icuff2	82.5	98.3	17.1	82.9	15.4
CAE-SRWC	76.68	78.5	**75**	**25**	53.5
SRCQW	**88.19**	98.09	**78.28**	**21.72**	**76.37**

Table 6.7: *Results by the SRCQW on the ISIC2019.*

	AC (%)	SP (%)	SE (%)	PR (%)	F1 (%)
128×128	70	56	84	65.6	73.7
192×192	74	72	76	73.1	74.5
256×192	70	76	64	72.7	68.1
256×256	74	68	80	71.4	75.5
512×512	**82**	**76**	**88**	**78.6**	**83**

ranking for melanoma classification reported in "Part 3: Disease Classification Task") that participated in the ISIC2017 challenge (N. C. Codella et al. 2018). Note that in Table 6.6, we use the institution name of these top ten competitors to indicate their methods (N. C. Codella et al. 2018), while SRCQW is the name of our proposed method. The SRCQW ranks first in terms of accuracy, sensitivity, miss rate, and Youden's index (J) and fourth according to specificity, while the CAE-SRWC ranks second in sensitivity and miss rate and third according to J-index. The Youden's index validates that the novel SRCQW method is best balanced among all NNs in diagnosing both benign and melanoma images through high rates in specificity and sensitivity. It is evident, from Table 6.6, that most of the methods provided results quite biased between benign and melanoma diagnosis (see SP, SE, and J in Table 6.6). In other words, they can diagnose almost every benign cases (high specificity) but fail to recognise most of melanoma cases (low sensitivity), especially MPG-UCIIIM ($J = 10.1\%$), icuff2 ($J = 15.4\%$), and icuff1 ($J = 16.1\%$). In contrast, the proposed SRCQW method diagnoses well not only the negative (benign) cases, but also the positive (malignant) cases having the highest $J = 76.37\%$.

In Table 6.7, we analyze the performance of the proposed method on ISIC2019 dataset according to the image sizes. It is evident that SRCQW with image size 512×512 obtains the best results in all metrics. More precisely, it make the substantial improvements up to 4% in accuracy and sensitivity, 5.5% in precision, and 7.5% according to F1 score.

Furthermore, to validate the advantages of the proposed SRCQW over NNs, we investigate the dependence of ResNet architectures (Pollastri et al. 2021) on the dataset size. Hence, three subsets with 10000, 5000, and 1000 samples are randomly selected from ISIC2019 dataset. One can see in Table 6.8 that, using only 1000 samples, the SRCQW method significantly outperforms all the ResNet NNs, which use for training away more images with every image sizes. In particular, with image size of 512×512, SRCQW using 1000 training images makes a substantial improvement of 4.7% compared to ResNet-152 using 10000 training images, which are huge gaps both in balanced accuracy and number of training images.

6.4 Conclusion

In this chapter, our primary contribution is the application of the novel SRWC, CAE-SRWC, and especially SRCQW methods for skin lesions classification. We applied the SRCQW method with different frequencies sub-bands. We validated its efficiency for skin lesion classification in the AQ, where the quaternions represent frequencies from the quaternion wavelet domain. More precisely, the novel SRCQW approach, which decomposes every image using the QWT, may implement any quaternion sub-band or pair of sub-bands to formulate the QWLasso problem. The latter is solved with the novel QFISTA method. Also, we determined that quaternions of low-frequency

Table 6.8: *Balanced Accuracy (columns 3, 4, 5) of NNs (Pollastri et al. 2021) and SRCQW on ISIC2019 with different Dataset Size (DS).*

	DS	512×512	256×256	128×128
ResNet-18	10000	69.9	69	62.5
ResNet-18	5000	61	61.7	56
ResNet-18	1000	43.1	45	41.6
ResNet-50	10000	75	72.2	61.4
ResNet-50	5000	66.3	63.7	51.1
ResNet-50	1000	45.6	46.4	41.3
ResNet-152	10000	77.3	73.9	64.8
ResNet-152	5000	68.9	63.3	57.2
ResNet-152	1000	52.2	49.5	43.3
SRCQW	1000	**82**	**74**	**70**

wavelet sub-bands provide a dictionary in the QW domain where the classification was conducted with highest accuracy compared with the other wavelet frequencies. This conclusion for lesion images confirms the one derived in (Ngo et al. 2018; W. Zou and Yan Li 2007) about human faces and 2D objects. SRWC and CAE-SRWC also show their promising results in skin lesion diagnosis, where SRWC is sensitive to melanoma detection and CAE-SRWC is well-balanced in classifying both melanoma and benign images.

An advantage of SRCQW over the NNs (Yu et al. 2018) is that the forward is well suited to provide very high classification statistics using middle size image datasets for training, while the NNs are able to exhibit their advantages when trained with very large datasets.

Given the above advantages, the novel SRCQW method meets the high expectation and demand for balanced and accurate skin lesion diagnosis (according to AC and J metrics in Table 6.3). Hence, it has the potential to be transferred to the clinical practice.

CHAPTER 7

Conclusion and future work

Chapter content

7.1 Introduction

This dissertation addressed the problem of exploiting the sparse representation (SR) in the transform domain, i.e. wavelet and quaternion wavelet (QW) domains. The present chapter summarizes the contributions of this study in the field of automated image classification with applications for faces (Extended YaleB, ARface, UMDAA-01), genders (AR gender), objects (COIL-100), digits (USPS, SVHN) or skin lesions (ISIC2017, ISIC2019) detection. In addition, it highlights some future directions to pursue.

More precisely, the proposed methods take advantages of SR and wavelet or QW domain in order to enhance the sparsity level of features and learn a simple and compact representation of the images. These advantages come from the fact that wavelets are naturally sparse and provide structural information about the image. Moreover, it helps to extract high discriminant features. Besides, QW has near shift-invariance and provides richer geometric information as well as higher sparsity of features than DWT.

SR was chosen as our main approach because it was recognized as a primary mechanism used in the early stages of visual cortex and considered as a main principle to efficiently represent complex data. SR has shown its efficiency in producing compact as well as simple representation of the images through only a small number of meaningful features. In addition, SR provide high robustness to noise, occlusion, and corruption in image classification tasks.

Table 7.1 summarizes a general comparison of the three proposed methods.

Table 7.1: *Comparison of the three proposed methods ((more ticks represent better performance).*

	SRWC	**SRCQW**	**CAE-SRWC**
Domain	Wavelet	Quaternion wavelet	Wavelet
Space	1D	4D	1D
Feature reduction	PCA	PCA	Autoencoder
Classification rule	Minimum residual	Minimum residual	Probability-based residual
Robust to training size	Yes	Yes	Yes
Computational cost	Low	High	Low
Faces classification	✓✓	✓✓✓	✓✓
Objects classification	✓✓	✓✓✓	✓✓
Skin lesions classification	✓	✓✓✓	✓✓

7.2 Summary and conclusions

A comprehensive literature review on image classification algorithms using recent techniques have been made in Chapter 2 before the three main contributions proposed in Chapter 3, 4 and 5. All the works are sparse representation (SR) based algorithms for image classification, where the first approach performs classification in the wavelet domain, while the second one classifies images in the quaternion wavelet domain, and the third one combines SR and neural network in the wavelet domain to enhance the classification performance.

More precisely:

In **Chapter 3**, our main contribution comes from the SR approach for classification using image features described by the low-frequency wavelet coefficients. In particular, an over-complete dictionary, which allows for representing a test sample from a given dataset, is built using the features generated by transforming the training samples into the wavelet domain. Then PCA is used to to reduce the dimension of the generated features and computational cost. Hence, a test sample can be represented as a sparse linear combination of base elements of the dictionary in the wavelet domain. This representation is naturally sparse, and help to reject test samples, which do not belong to

the dataset (Wright, A. Y. Yang, et al. 2008). Moreover, the wavelets promote sparsity and provide structural information about the image, which boosts the classification performance. To validate the capabilities and underline the advantages of the novel SRWC, we conducted an extensive number of experiments using publicly available datasets including faces and object. By comparing our results with others, we prove that the proposed SRWC outperforms some state-of-the-art methods in the field.

In **Chapter 4**, in order to improve the classification performance of the SRWC, our primary contribution is the development of the novel and robust sparsity-inducing SRCQW method for image classification in the algebra of quaternions. We introduce the novel method for multi-class image classification based on the SR approach, which operates in the quaternion wavelet domain. The advantage of this domain is its near shift invariance which is not the case for the DWT used in the SRWC proposed in chapter 3. In this method, we only make use of the image features described by the information from the low-frequency coefficients of quaternion wavelet, which represent the most important components of the image in the coarsest level, to construct the sparse dictionary and the classifier in the 4D space of quaternions. The sparse quaternion dictionary is constructed by the Quaternion Wavelet coefficients in the low-frequency sub-bands of the training samples. To estimate the quaternion SR vector in the sparse coding stage, we formulate the QWLasso model using quaternion l_1 minimization. In order to solve the QWLasso minimization model and determine the quaternion SR vector, we develop the novel QFISTA. In particular, we develop in the novel QFISTA an upper bound for the QWLasso model and use the upper bound as an approximation that establishes the iterative scheme to find the quaternion SR vector. The fusion of the wavelets and the SR model in the quaternion wavelet domain makes the novel QWLasso method achieve high accuracy of classification. To the best of our knowledge, the proposed SRCQW is the first approach that uses information from the quaternion wavelet domain to solve the minimization problem QWLasso for classification in the 4D space of the quaternion algenbra. Our experimental validation was conducted on several public datasets, which consist of faces, genders, and objects. The experimental results show that the proposed method yields substantial accuracy improvement over the state-of-the-art methods in the field including Neural Network based approaches.

In **Chapter 5**, we propose to use SR as a layer of a neural network. In particular, a convolutional autoencoder architecture including a sparse latent layer is constructed in

the wavelet domain. Analogous to the work of SRWC method, only image low-frequency wavelet sub-bands are utilized as the input of the network. To assign identity to the unlabeled samples, a residual-based probabilistic criterion is exploited based on the estimated sparse coefficients. Extensive experiments conducted on several public datasets, including faces, genders, objects, and digits, validated the superior of the proposed methods over various recent neurals networks.

Finally, **Chapter 6** applies the proposed methods to skin lesion image classification. Up until now, it is always challenging to automatically detect melanoma images from benign images due to various barriers. This leads to our primary contribution in this chapter, which is the application of the novel SRWC, SRCQW, and CAE-SRWC methods. Further, we investigate the application of the SR based approach with low, high, and mixed quaternion wavelet frequencies. Using the public skin lesion image dataset ISIC2017 and ISIC2019, we experimentally determined that creating dictionary with low-frequency wavelet sub-bands leads to the most accurate classification of melanoma and benign skin lesions. To validate the capabilities of the novel approaches, we compared them with multiple contemporary methods including neural networks. While SRWC is sensitive to melanoma detection (sensitivity=82.06%), SRCQW and CAE-SRWC methods meet the high expectation and demand for balanced and accurate skin lesion diagnosis. Hence, they have the potential to be transferred to the clinical practice

7.3 Future works and perspectives

In the following, we conclude by mentioning/listing/detailing some of the possible extensions/directions/ideas to be investigated inspired by the achievements in this dissertation.

Dictionary Learning: A key point that we have intentionally paid less attention to, in this study, is the necessity to learn the dictionaries in parallel with the sparse codes update. Hence, the learned dictionaries and the obtained sparse codes will be concurrently optimized.

Data labels: In this dissertation, data labels do not play an important role in our approaches. As proven in the work by (Jiang et al. 2013), labels need to be considered in further works to enhance the performance of classification.

Sparse representation and Kernel: Kernel, a familiar technique in machine

learning, can be fused with sparse representation to better discriminate the input samples. In particularly, the input samples are implicitly mapped into a high-dimensional space, namely kernel feature space, with the help of a nonlinear kernel function. This combination promisingly increases the accuracy of classification, because it discriminates the different samples from different classes.

Sparse representation and Deep networks: We can not stay out of the deep revolution. Taking advantage of both sparse representation and deep neural networks, we can combine the best of both strategies to construct multilayer sparse coding networks or sparse deep neural networks. These networks are theoretically expected to achieve outstanding enhancements over their individual counterparts, which can be seen clearly in Chapter 5 with the proposed CAE-SRWC method. Hence, it is worth investigating more on the fusion of the two methodologies for the future work.

Deployment to Practice: The application of the proposed methods on skin lesion image demonstrate their potential to be deployed to clinical practice. However, they need improving in some aspects to perfectly fit the clinical demands.

Implementation: We can improve the running time of the proposed methods by implementing them with the help of powerful and efficient GPUs.

References

Abavisani, Mahdi and Vishal M Patel (2019). "Deep sparse representation-based classification". In: *IEEE Signal Processing Letters* 26.6, pp. 948–952 *Cited on pages 74, 75, 82.*

Abuzaghleh, Omar, Buket D Barkana, and Miad Faezipour (2014). "Automated skin lesion analysis based on color and shape geometry feature set for melanoma early detection and prevention". In: *IEEE Long Island Systems, Applications and Technology (LISAT) Conf.* IEEE. May, Farmingdale, New York, USA, pp. 1–6 *Cited on page 90.*

Aggarwa, Charu C (2014). *Data Classification: Algorithms and Applications.* 1st edition. Chapman & Hall/CRC *Cited on pages 2, 3.*

Aharon, Michal, Michael Elad, and Alfred Bruckstein (2006a). "k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation". In: *IEEE Trans. on Signal Processing* 54.11, pp. 4311–4322 *Cited on page 28.*

— (2006b). "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation". In: *IEEE Trans. on Signal Processing (TSP)* 54.11, pp. 4311–4322 *Cited on pages 6, 12, 16, 26.*

Alom, Md Zahangir, Theodore Josue, Md Nayim Rahman, Will Mitchell, Chris Yakopcic, and Tarek M Taha (2018). "Deep versus wide convolutional neural networks for object recognition on neuromorphic system". In: *IEEE Int. Joint Conf. on Neural Networks (IJCNN).* July, Rio, Brazil, pp. 1–8 *Cited on pages 62, 65, 66.*

Anthimopoulos, Marios, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou (2016). "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network". In: *IEEE Trans. on Medical Imaging* 35.5, pp. 1207–1216 *Cited on page 2.*

Astudillo, Natasha M, Reginald Bolman, and Nikolay M Sirakov (2020). "Classification with Stochastic Learning Methods and Convolutional Neural Networks". In: *SN Computer Science* 1.3, pp. 1–9 *Cited on page 91.*

Balakrishnama, Suresh and Aravind Ganapathiraju (1998). "Linear discriminant analysis-a brief tutorial". In: *Institute for Signal and information Processing* 18.1998, pp. 1–8 *Cited on page 4.*

Balch, Charles M, Antonio C Buzaid, Seng-Jaw Soong, Michael B Atkins, Natale Cascinelli, Daniel G Coit, Irvin D Fleming, Jeffrey E Gershenwald, Alan Houghton Jr, M Kirkwood John M andK McMasters, M.F. Mihm, D.L. Morton, D.S. Reintgen, M.I. Ross, A Sober, J.A. Thompson, and J.F. Thompson (2001). "Final ver. of the American Joint Committee on Cancer staging system for cutaneous melanoma". In: *J. of Clinical Oncology* 19.16, pp. 3635–3648 *Cited on page 87.*

Ballerini, Lucia, Robert B Fisher, Ben Aldridge, and Jonathan Rees (2013). "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions". In: *Color Medical Image Analysis*, pp. 63–86 *Cited on pages 89, 90.*

Barata, Catarina, Margarida Ruela, Mariana Francisco, Teresa Mendonça, and Jorge S Marques (2013). "Two systems for the detection of melanomas in dermoscopy images using texture and color features". In: *IEEE Systems Journal* 8.3, pp. 965–979 *Cited on pages 90, 91.*

Basri, Ronen and David W Jacobs (2003). "Lambertian reflectance and linear subspaces". In: *IEEE TPAMI* 25.2, pp. 218–233 *Cited on page 32.*

Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool (2008). "Speeded-up robust features (SURF)". In: *Computer vision and image understanding* 110.3, pp. 346–359 *Cited on page 3.*

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool (2006). "Surf: Speeded up robust features". In: *European Conf. on Computer Vision*. Springer. May, Graz, Austria, pp. 404–417 *Cited on page 3.*

Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM J. on Imaging Sciences* 2.1, pp. 183–202 *Cited on pages 34, 46, 57, 59, 70.*

Binder, Michael, Margot Schwarz, Alexander Winkler, Andreas Steiner, Alexandra Kaider, Klaus Wolff, and Hubert Pehamberger (1995). "Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists". In: *Archives of dermatology* 131.3, pp. 286–291 *Cited on page 88.*

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. 1st edition. Springer-Verlag New York *Cited on pages 2, 3.*

Bosch, Anna, Andrew Zisserman, and Xavier Munoz (2007). "Image classification using random forests and ferns". In: *2007 IEEE 11th Int. Conf. on Computer Vision*. Ieee. October, Rio de Janeiro, Brasil, pp. 1–8 *Cited on page 4.*

Boureau, Y-Lan, Jean Ponce, and Yann LeCun (2010). "A theoretical analysis of feature pooling in visual recognition". In: *Proceedings of the 27th Int. Conf. on Machine Learning (ICML-10)*. June, Haifa, Israel, pp. 111–118 *Cited on page 21.*

Bourlard, Hervé and Yves Kamp (1988). "Auto-association by multilayer perceptrons and singular value decomposition". In: *Biological Cybernetics* 59.4, pp. 291–294 *Cited on page 22.*

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine Learning* 3.1, pp. 1–122 *Cited on pages 12, 19.*

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal (2018). "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer J. for Clinicians* 68.6, pp. 394–424 *Cited on page 2.*

Bruna, Joan and Stéphane Mallat (2013). "Invariant scattering convolution networks". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35.8, pp. 1872–1886 *Cited on page 23.*

Bulow, Thomas (1999). "Hypercomplex spectral signal representations for the processing and analysis of images". In: *Ph.D thesis, Christian-Albrechts-Universitat zu Kiel* *Cited on pages 45, 50.*

Candès, Emmanuel J, Justin Romberg, and Terence Tao (2006). "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Trans. on information theory* 52.2, pp. 489–509 *Cited on page 26.*

Chan, Tsung-Han, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma (2015). "PCANet: A simple deep learning baseline for image classification?" In: *IEEE Trans. on Image Processing (TIP)* 24.12, pp. 5017–5032 *Cited on pages 4, 62, 65.*

Chan, Wai Lam, Hyeokho Choi, and Richard G Baraniuk (2004). "Quaternion wavelets for image analysis and processing". In: *IEEE ICIP*. Vol. 5. October, Singapore, pp. 3057–3060 *Cited on pages 45, 50, 51.*

— (2008). "Coherent multiscale image processing using dual-tree quaternion wavelets". In: *IEEE Trans. on Image Processing* 17.7, pp. 1069–1082 *Cited on pages 45, 50, 51.*

Chapelle, Olivier, Patrick Haffner, and Vladimir N Vapnik (1999). "Support vector machines for histogram-based image classification". In: *IEEE Trans. on Neural Networks* 10.5, pp. 1055–1064 *Cited on page 4.*

Chen, Yi, Nasser M Nasrabadi, and Trac D Tran (2013). "Hyperspectral image classification via kernel sparse representation". In: *IEEE Trans. on Geoscience and Remote Sensing* 51.1, pp. 217–231 *Cited on pages 6, 26, 28.*

Cheng, Heng-Da, Xiaopeng Cai, Xiaowei Chen, Liming Hu, and Xueling Lou (2003). "Computer-aided detection and classification of microcalcifications in mammograms: a survey". In: *Pattern Recognition* 36.12, pp. 2967–2991 *Cited on page 2.*

Cheng, Hong, Zicheng Liu, Lu Yang, and Xuewen Chen (2013). "Sparse representation and learning in visual recognition: Theory and applications". In: *Signal Processing* 93.6, pp. 1408–1425 *Cited on page 4.*

Choromanska, Anna, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun (2015). "The loss surfaces of multilayer networks". In: *Artificial Intelligence and Statistics*. May, San Diego, CA, USA, pp. 192–204 *Cited on page 24.*

Ciresan, Dan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber (2012). "Deep neural networks segment neuronal membranes in electron microscopy images". In: *Advances in Neural Information Processing Systems* 25, pp. 2843–2851 *Cited on page 22.*

Codella, Noel, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith (2015). "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images". In: *Int. Workshop on Machine Learning in Medical Imaging*. Springer. October 5-9, Munich, Germany, pp. 118–126 *Cited on pages 2, 91.*

Codella, Noel CF, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern (2018). "Skin lesion analysis toward melanoma detection: A challenge at the 2017 Int. symposium on biomedical imaging (isbi), hosted by the Int. skin imaging collaboration (isic)". In: *2018 IEEE 15th Int. Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. April, Washington DC, USA, pp. 168–172 *Cited on pages 90–93, 98.*

Combalia, Marc, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, and Josep Malvehy (2019). "BCN20000: Dermoscopic lesions in the wild". In: *arXiv:1908.02288* *Cited on pages 90, 92, 93.*

Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE CVPR*. Vol. 1. IEEE. June, San Diego, CA, USA, pp. 886–893 *Cited on page 3.*

Dao, Minh, Nam H Nguyen, Nasser M Nasrabadi, and Trac D Tran (2016). "Collaborative multi-sensor classification via sparsity-based representation". In: *IEEE Trans. on Signal Processing* 64.9, pp. 2400–2415 *Cited on page 6.*

Defossez, Gautier, Sandra Le Guyader-Peyrou, Zoé Uhry, Pascale Grosclaude, Marc Colonna, Emmanuelle Dantony, Patricia Delafosse, Florence Molinié, Anne-Sophie Woronoff, and Anne-Marie Bouvier (2019). "Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018". In: *Etude à partir des registres des cancers du réseau Francim. Résultats préliminaires. Synthèse. Saint-Maurice (Fra): Santé Publique France* *Cited on page 2.*

Donoho, David L (2006). "Compressed sensing". In: *IEEE Trans. on information theory* 52.4, pp. 1289–1306 *Cited on pages 6, 26.*

Dubitzky, Werner, Martin Granzow, and Daniel P Berrar (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media *Cited on pages 36, 62, 92.*

Duda, Richard O, Peter E Hart, and David G Stork (2001). *Pattern classification*. 2nd edition. John Willey & Sons *Cited on pages 2, 3.*

Elad, Michael (2010). *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media *Cited on pages 4, 5, 26, 36, 44, 74, 75.*

Elad, Michael and Michal Aharon (2006). "Image denoising via sparse and redundant representations over learned dictionaries". In: *IEEE Trans. on Image Processing* 15.12, pp. 3736–3745 *Cited on pages 19, 26.*

Elad, Michael, Mario AT Figueiredo, and Yi Ma (2010). "On the role of sparse and redundant representations in image processing". In: *Proceedings of the IEEE* 98.6, pp. 972–982 *Cited on pages 6, 26.*

Elhamifar, Ehsan and René Vidal (2011). "Robust classification using structured sparse representation". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE. June, Colorado, USA, pp. 1873–1879 *Cited on page 34.*

Fawzi, Alhussein (2016). "Robust Image Classification: Analysis and Applications". PhD thesis. Ecole Polytechnique Fédérale de Lausanne *Cited on page 21.*

Fawzi, Alhussein, Mike Davies, and Pascal Frossard (2015). "Dictionary learning for fast classification based on soft-thresholding". In: *Int. J. of Computer Vision* 114.2-3, pp. 306–321 *Cited on page 21.*

Fukushima, Kunihiko (1979). "Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron". In: *IEICE Technical Report, A* 62.10, pp. 658–665 *Cited on page 22.*

Ganster, Harald, P Pinz, Reinhard Rohrer, Ernst Wildling, Michael Binder, and Harald Kittler (2001). "Automated melanoma recognition". In: *IEEE Trans. on Medical Imaging* 20.3, pp. 233–239 *Cited on pages 89, 90.*

Georghiades, Athinodoros S., Peter N. Belhumeur, and David J. Kriegman (2001). "From few to many: Illumination cone models for face recognition under variable lighting and pose". In: *IEEE TPAMI* 23.6, pp. 643–660 *Cited on pages 8, 9, 36, 37, 62, 64, 82, 83, 85.*

Georgiou, Theodoros, Yu Liu, Wei Chen, and Michael Lew (2020). "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision". In: *Int. J. of Multimedia Information Retrieval* 9.3, pp. 135–170 *Cited on pages 4, 20, 23.*

Ghazali, Kamarul Hawari, Mohd Fais Mansor, Mohd Marzuki Mustafa, and Aini Hussain (2007). "Feature extraction technique using discrete wavelet transform for image classification". In: *2007 5th Stud. Conf. on Research and Development*. IEEE. December, Selangor, Malaysia, pp. 1–4 *Cited on page 28.*

Girard, Patrick R (2007). *Quaternions, Clifford algebras and relativistic physics*. Springer Science & Business Media *Cited on pages 48, 49.*

Giryes, Raja, Guillermo Sapiro, and Alex M Bronstein (2016). "Deep neural networks with random gaussian weights: A universal classification strategy?" In: *IEEE Trans. on Signal Processing* 64.13, pp. 3444–3457 *Cited on pages 23, 24.*

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep sparse rectifier neural networks". In: *Proceedings of the 14th Int. Conf. on Artificial Intelligence and Statistics*. April, Lauderdale, FL, USA, pp. 315–323 *Cited on page 21.*

Gonzalez-Diaz, Ivan (2018). "Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis". In: *IEEE J. of Biomedical and Health Informatics* 23.2, pp. 547–559 *Cited on page 2.*

Guo, Tiantong, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga (2017). "Deep wavelet prediction for image super-resolution". In: *The IEEE CVPR Workshops*. July, Honolulu, HI, USA *Cited on pages 30, 31.*

Hamilton, William Rowan (1844). "LXXVIII. On quaternions; or on a new system of imaginaries in Algebra: To the editors of the Philosophical Magazine and Journal". In: *The London, Edinburgh, Dublin* 25.169, pp. 489–495 *Cited on page 48.*

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*. June, Las Vegas, NV, USA, pp. 770–778 *Cited on pages 75, 76.*

Hearst, Marti A., Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf (1998). "Support vector machines". In: *IEEE Intelligent Systems and Their Applications* 13.4, pp. 18–28 *Cited on page 4.*

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh (2006). "A fast learning algorithm for deep belief nets". In: *Neural Computation* 18.7, pp. 1527–1554 *Cited on pages 4, 22.*

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786, pp. 504–507 *Cited on pages 4, 22, 23.*

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5, pp. 359–366 *Cited on page 4.*

Hoyer, Patrik O (2004). "Non-negative matrix factorization with sparseness constraints". In: *J. of Machine Learning Research* 5.Nov, pp. 1457–1469 *Cited on pages 40, 68.*

Huang, Xin, Liangpei Zhang, and Pingxiang Li (2008). "A multiscale feature fusion approach for classification of very high resolution satellite imagery based on wavelet transform". In: *Int. J. of Remote Sensing* 29.20, pp. 5923–5941 *Cited on page 28.*

Hull, Jonathan J. (1994). "A database for handwritten text recognition research". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 16.5, pp. 550–554 *Cited on pages 9, 82, 85.*

Jaderberg, Max, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu (2015). "Spatial transformer networks". In: *Advances in Neural Information Processing Systems*. December, Montreal, Quebec, Canada, pp. 2017–2025 *Cited on page 20.*

Jaworek-Korjakowska, Joanna (2012). "Automatic detection of melanomas: an application based on the ABCD criteria". In: *Information Technologies in Biomedicine*, pp. 67–76 *Cited on page 90.*

Jerant, Anthony, Jennifer Johnson, Catherine Demastes Sheridan, and Timothy Caffrey (2000). "Early detection and treatment of skin cancer." In: *American family physician* 62.2 *Cited on page 87.*

Jian, Muwei and Lei Liu (2009). "Texture image classification using visual perceptual texture features and gabor wavelet". In: *J. of Computers* 4.8, p. 763 *Cited on page 29.*

Jiang, Zhuolin, Zhe Lin, and Larry S Davis (2013). "Label consistent K-SVD: Learning a discriminative dictionary for recognition". In: *IEEE TPAMI* 35.11, pp. 2651–2664 *Cited on pages 6, 12, 15, 16, 26–28, 35, 36, 42, 44, 45, 62, 63, 82, 104.*

Jolliffe, Ian (2011). *Principal component analysis*. Springer *Cited on page 55.*

Jolliffe, Ian T (1986). "Principal Component Analysis and Factor Analysis". In: *Principal Component Analysis*. Springer, pp. 115–128 *Cited on pages 4, 31.*

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* *Cited on page 21.*

Korotkov, Konstantin and Rafael Garcia (2012). "Computerized analysis of pigmented skin lesions: a review". In: *Artificial intelligence in medicine* 56.2, pp. 69–90 *Cited on page 90.*

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90 *Cited on pages 4, 19, 20, 22.*

Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce (2006). "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *2006 IEEE CVPR*. Vol. 2. IEEE. June, New York, USA, pp. 2169–2178 *Cited on page 4.*

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 *Cited on pages 20, 22, 23.*

LeCun, Yann A, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller (2012). "Efficient backprop". In: *Neural Networks: Tricks of the Trade*. Springer, pp. 9–48 *Cited on page 21.*

Li, Feng, JM Zuraday, and Wei Wu (2018). "Sparse Representation Learning of Data by Autoencoder with $L_{1/2}$ Regularization". In: *Neural Network World* 28.2, pp. 133–147 *Cited on page 75.*

Li, Sheng and Yun Fu (2016). "Learning robust and discriminative subspace with low-rank constraints". In: *IEEE Trans. on neural networks and learning systems* 27.11, pp. 2160–2173 *Cited on page 37.*

Li, Wei and Qian Du (2014). "Gabor-filtering-based nearest regularized subspace for hyperspectral image classification". In: *IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing* 7.4, pp. 1012–1022 *Cited on page 4.*

Li, Yuexiang and Linlin Shen (2018). "Skin lesion analysis towards melanoma detection using deep learning network". In: *Sensors* 18.2 *Cited on pages 4, 91–93, 96.*

Liang, Xinyue, Alireza M Javid, Mikael Skoglund, and Saikat Chatterjee (2018). "Distributed Large Neural Network with Centralized Equivalence". In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. April, Calgary, AB, Canada, pp. 2976–2980 *Cited on pages 62, 65.*

Livni, Roi, Shai Shalev-Shwartz, and Ohad Shamir (2014). "On the computational efficiency of training neural networks". In: *Advances in Neural Information Processing Systems*. December, Montreal, Quebec, Canada, pp. 855–863 *Cited on page 24.*

Lowe, David G (2004). "Distinctive image features from scale-invariant keypoints". In: *Int. J. of computer vision* 60.2, pp. 91–110 *Cited on page 3.*

Lu, Xiaoqiang and Xuelong Li (2014). "Group sparse reconstruction for image segmentation". In: *Neurocomputing* 136, pp. 41–48 *Cited on pages 26, 75.*

Ma, Li, Melba M Crawford, and Jinwen Tian (2010). "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification". In: *IEEE Trans. on Geoscience and Remote Sensing* 48.11, pp. 4099–4109 *Cited on page 4.*

Maas, Andrew L, Awni Y Hannun, and Andrew Y Ng (2013). "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. ICML*. Vol. 30. 1. June, Atlanta, USA, pp. 3–8 *Cited on page 21.*

Mahendran, Aravindh and Andrea Vedaldi (2015). "Understanding deep image representations by inverting them". In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*. June, Boston, MA, USA, pp. 5188–5196 *Cited on page 22.*

Mairal, Julien, Francis Bach, and Jean Ponce (2014). "Sparse modeling for image and vision processing". In: *arXiv preprint arXiv:1411.3230* *Cited on page 19.*

Mairal, Julien, Michael Elad, and Guillermo Sapiro (2007). "Sparse representation for color image restoration". In: *IEEE Trans. on Image Processing* 17.1, pp. 53–69 *Cited on page 18.*

Mallat, Stephane (2008). *A wavelet tour of signal processing: the sparse way*. Academic press *Cited on pages 30, 31.*

Martens, James (2010). "Deep learning via hessian-free optimization." In: *ICML*. Vol. 27. June, Haifa, Israel, pp. 735–742 *Cited on page 21.*

Martens, James and Roger Grosse (2015). "Optimizing neural networks with kronecker-factored approximate curvature". In: *Int. Conf. on Machine Learning*. July, Lille, France, pp. 2408–2417 *Cited on page 21.*

Martinez, Aleix M (1998). "The AR face database". In: *CVC Technical Report24* *Cited on pages 8, 9, 36–38, 62–64, 82–85.*

Mete, Mutlu and Nikolay Metodiev Sirakov (2012). "Dermoscopic diagnosis of melanoma in a 4D space constructed by active contour extracted features". In: *CMIG Journal* 36.7, pp. 572–579 *Cited on pages 89, 90.*

Mishra, Nabin K and M Emre Celebi (2016). "An overview of melanoma detection in dermoscopy images using image processing and machine learning". In: *arXiv preprint arXiv:1601.07843* *Cited on page 2.*

Molinaro, Annette M, Richard Simon, and Ruth Pfeiffer (2005). "Prediction error estimation: a comparison of resampling methods". In: *Bioinformatics* 21.15, pp. 3301–3307 *Cited on pages 36, 62, 92.*

Moradi, Nooshin and Nezam Mahdavi-Amiri (2019). "Kernel sparse representation based model for skin lesions segmentation and classification". In: *Computer methods and programs in biomedicine* 182 *Cited on pages 26, 91.*

Nalband, Saif, Aditya Sundar, A Amalin Prince, and Anita Agarwal (2016). "Feature selection and classification methodology for the detection of knee-joint disorders". In: *Computer Methods and Programs in Biomedicine* 127, pp. 94–104 *Cited on page 2.*

Nene, Sameer A, Shree K Nayar, and Hiroshi Murase (1996). "Columbia Object Image Library (COIL-100)". In: *Technical Report CUCS-006-96* *Cited on pages 8, 9, 37, 62–64, 82, 83, 85.*

Netzer, Yuval, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng (2011). "Reading digits in natural images with unsupervised feature learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning* *Cited on pages 9, 74, 82, 85.*

Ngo, Long H, Marie Luong, Nikolay M Sirakov, Thuong Le-Tien, Sebastien Guerif, and Emmanuel Viennet (2018). "Sparse Representation Wavelet Based Classification". In: *25th IEEE Int. Conf. on Image Processing (ICIP)*. IEEE. October, Athens, Greece, pp. 2974–2978 *Cited on pages 4, 45, 46, 54, 62, 65, 69, 74, 79, 82, 84, 97, 100.*

Nguyen, Tan-Sy, Long H Ngo, Marie Luong, Mounir Kaaniche, and Azeddine Beghdadi (2020). "Convolution Autoencoder-Based Sparse Representation Wavelet for Image Classification". In: *2020 IEEE 22nd Int. Workshop on Multimedia Signal Processing (MMSP)*. IEEE. September, Tampere, Finland, pp. 1–6 *Cited on pages 92, 97.*

Noroozi, Navid and Ali Zakerolhosseini (2016). "Computer assisted diagnosis of basal cell carcinoma using Z-transform features". In: *J. of Visual Communication and Image Representation* 40, pp. 128–148 *Cited on page 91.*

Olshausen, Bruno A (2003). "Principles of image representation in visual cortex". In: *The visual Neurosciences* 2, pp. 1603–1615 *Cited on page 44.*

Olshausen, Bruno A and David J Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583, pp. 607–609 *Cited on pages 4, 6, 26, 44, 74.*

— (1997). "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* 37.23, pp. 3311–3325 *Cited on pages 6, 26, 44.*

Papyan, Vardan, Yaniv Romano, and Michael Elad (2017). "Convolutional neural networks analyzed via convolutional sparse coding". In: *The J. of Machine Learning Research* 18.1, pp. 2887–2938 *Cited on page 74.*

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman (2015). "Deep face recognition". In: *Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, Proceedings of the British Machine Vision Conf. (BMVC)*. BMVA Press. September, Swansea, UK, pp. 41.1–41.12 *Cited on page 20.*

Pati, Yagyensh Chandra, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad (1993). "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition". In: *Proceedings of 27th Asilomar Conf. on Signals, Systems and Computers*. IEEE. November, Pacific Grove, CA, USA, pp. 40–44 *Cited on page 18.*

Pollastri, Federico, Mario Parreño, Juan Maroñas, Federico Bolelli, Roberto Paredes, Daniel Ramos, and Costantino Grana (2021). "A Deep Analysis on High Resolution Dermoscopic Image Classification". In: *IET Research J.* ISSN: 1751-8644 *Cited on page 99.*

Quinlan, J. Ross (1986). "Induction of decision trees". In: *Machine Learning* 1.1, pp. 81–106 *Cited on page 4.*

Ranzato, Marc, Christopher Poultney, Sumit Chopra, and Yann LeCun (2007). "Efficient learning of sparse representations with an energy-based model". In: *Advances in Neural Information Processing Systems* 19, p. 1137 *Cited on page 23.*

Rastgoo, Mojdeh, Rafael Garcia, Olivier Morel, and Franck Marzani (2015). "Automatic differentiation of melanoma from dysplastic nevi". In: *CMIG Journal* 43, pp. 44–52 *Cited on pages 90, 91.*

Rastgoo, Mojdeh, Guillaume Lemaitre, Olivier Morel, Joan Massich, Rafael Garcia, Fabrice Meriaudeau, Franck Marzani, and Désiré Sidibé (2016). "Classification of melanoma lesions using sparse coded features and random forests". In: *Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis*. Int. Society for Optics and Photonics *Cited on pages 89, 91.*

Rebouças Filho, Pedro Pedrosa, Solon Alves Peixoto, Raul Victor Medeiros da Nóbrega, D Jude Hemanth, Aldisio Gonçalves Medeiros, Arun Kumar Sangaiah, and Victor Hugo C

de Albuquerque (2018). “Automatic histologically-closer classification of skin lesions”. In: *CMIG Journal* 68, pp. 40–54 *Cited on pages 92, 96, 97.*

Riaz, Farhan, Ali Hassan, Muhammad Younis Javed, and Miguel Tavares Coimbra (2014). “Detecting melanoma in dermoscopy images using scale adaptive local binary patterns”. In: *36th Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society*. IEEE. August 26-30, Chicago, Illinois, USA, pp. 6758–6761 *Cited on pages 90, 91.*

Ruela, Margarida, Catarina Barata, and Jorge S Marques (2013). “What is the role of color symmetry in the detection of melanomas?” In: *Int. Symposium on Visual Computing*. Springer. July 29-31, Rethymnon, Crete, Greece, pp. 1–10 *Cited on page 90.*

Ruela, Margarida, Catarina Barata, Teresa Mendonça, and Jorge S Marques (2013). “On the role of shape in the detection of melanomas”. In: *2013 8th Int. Sym. on Image and Signal Processing and Analysis (ISPA)*. IEEE. September 4-6, Trieste, Italy, pp. 268–273 *Cited on pages 90, 91.*

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal (2019). “Cancer statistics, 2019”. In: *CA: a cancer J. for clinicians* 69.1, pp. 7–34 *Cited on page 87.*

Silveira, Margarida, Jacinto C Nascimento, Jorge S Marques, André RS Marçal, Teresa Mendonça, Syogo Yamauchi, Junji Maeda, and Jorge Rozeira (2009). “Comparison of segmentation methods for melanoma diagnosis in dermoscopy images”. In: *IEEE J. Selected Topics in Signal Processing* 3.1, pp. 35–45 *Cited on page 88.*

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* *Cited on page 22.*

Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* *Cited on pages 20, 22, 75, 76.*

Sirakov, Nikolay Metodiev, Mutlu Mete, Richard Selvaggi, and Marie Luong (2015). “New accurate automated melanoma diagnosing systems”. In: *2015 Int. Conf. on Healthcare Informatics*. IEEE. October 21-23, Dallas, TX, USA, pp. 374–379 *Cited on pages 88–90.*

Sirakov, Nikolay Metodiev, Ye-Lin Ou, and Mutlu Mete (2015). “Skin lesion feature vectors classification in models of a Riemannian manifold”. In: *Annals of Mathematics and Artificial Intelligence* 75.1-2, pp. 217–229 *Cited on pages 89, 90.*

Smolensky, Paul (1986). “Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory”. In: *MIT Press, Cambridge, MA, USA* 15 *Cited on page 22.*

Soulard, Raphaël and Philippe Carré (2011). “Quaternionic wavelets for texture classification”. In: *Pattern Recognition Letters* 32.13, pp. 1669–1678 *Cited on pages 45, 46.*

Sousa, Rafael Teixeira and Larissa Vasconcellos de Moraes (2017). "Araguaia medical vision lab at ISIC 2017 skin lesion classification challenge". In: *arXiv preprint arXiv:1703.00856* *Cited on pages 91, 92, 96.*

Spratling, Michael W (2013). "Image segmentation using a sparse coding model of cortical area V1". In: *IEEE TIP* 22.4, pp. 1631–1643 *Cited on pages 6, 26.*

Srinivas, Umamahesh, Hojjat Seyed Mousavi, Vishal Monga, Arthur Hattel, and Bhushan Jayarao (2014). "Simultaneous sparsity model for histopathological image representation and classification". In: *IEEE Trans. on Medical Imaging* 33.5, pp. 1163–1179 *Cited on page 6.*

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The J. of Machine Learning Research* 15.1, pp. 1929–1958 *Cited on page 22.*

Subakan, Özlem N and Baba C Vemuri (2011). "A quaternion framework for color image smoothing and segmentation". In: *Int. J. of Computer Vision* 91.3, pp. 233–250 *Cited on page 45.*

Suzuki, Kenji (2013). "Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey". In: *IEICE Trans. on Information and Systems* 96.4, pp. 772–783 *Cited on page 2.*

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI Conf. on Artificial Intelligence*. Vol. 31. 1. February, San Francisco, CA, USA *Cited on page 20.*

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015). "Going deeper with convolutions". In: *Proc. of the IEEE CVPR*. June, Boston, MA, USA, pp. 1–9 *Cited on page 20.*

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*. June, Las Vegas, NV, USA, pp. 2818–2826 *Cited on pages 75, 76.*

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf (2014). "Deepface: Closing the gap to human-level performance in face verification". In: *Proc. of the IEEE CVPR*. June, Columbus, OH, USA, pp. 1701–1708 *Cited on page 20.*

Tian, Chunwei, Qi Zhang, Guanglu Sun, Zhichao Song, and Siyan Li (2018). "FFT consolidated sparse and collaborative representation for image classification". In: *Arabian J. for Science and Engineering* 43.2, pp. 741–758 *Cited on page 29.*

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *J. of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288 *Cited on pages 13, 14, 34.*

Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler (2018). "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5 *Cited on pages 90, 92, 93.*

Unser, Michael (1995). "Texture classification and segmentation using wavelet frames". In: *IEEE Trans. on Image Processing* 4.11, pp. 1549–1560 *Cited on page 26.*

Vedaldi, Andrea and Karel Lenc (2015). "Matconvnet: Convolutional neural networks for matlab". In: *Proceedings of the 23rd ACM Int Conf. on Multimedia*. October, Brisbane, Australia, pp. 689–692 *Cited on pages 20, 21.*

Verma, Luxmi, Sangeet Srivastava, and PC Negi (2016). "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data". In: *J. of Medical Systems* 40.7, pp. 1–7 *Cited on page 2.*

Vestergaard, ME, PHPM Macaskill, PE Holt, and SW Menzies (2008). "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting". In: *British J. Dermatology* 159.3, pp. 669–676 *Cited on page 88.*

Vu, Tiep Huu and Vishal Monga (2017). "Fast Low-Rank Shared Dictionary Learning for Image Classification". In: *IEEE TIP* 26.11, pp. 5160–5175 *Cited on pages 12, 34, 62, 63.*

Wager, Stefan, William Fithian, Sida Wang, and Percy S Liang (2014). "Altitude training: Strong bounds for single-layer dropout". In: *Advances in Neural Information Processing Systems*. December, Montreal, Quebec, Canada, pp. 100–108 *Cited on page 22.*

Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology *Cited on page 20.*

Wan, Li, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus (2013). "Regularization of neural networks using dropconnect". In: *Int. Conf. on Machine Learning*. June, Atlanta, GA, USA, pp. 1058–1066 *Cited on page 22.*

Wei, Jiang-Shu, Jian-Cheng Lv, and Chun-Zhi Xie (2016). "A new sparse representation classifier (SRC) based on probability judgement rule". In: *2016 Int. Conf. on Information System and Artificial Intelligence (ISAI)*. IEEE. June, Hong Kong, China, pp. 338–342 *Cited on pages 75, 80.*

Wright, John, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan (2010). "Sparse representation for computer vision and pattern recognition". In: *Proceedings of the IEEE* 98.6, pp. 1031–1044 *Cited on pages 6, 26, 38.*

Wright, John, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma (2008). "Robust face recognition via sparse representation". In: *IEEE TPAMI* 31.2, pp. 210–227 *Cited on pages 6, 12, 14, 26–28, 32–36, 39, 41, 44, 45, 57, 62, 63, 69, 74–76, 80, 103.*

Xu, YH, YR Zhao, WX Hong, et al. (2010). "Based on quaternion of small amplitude phase representation and block voting strategy approach for face recognition". In: *Application Research of Computers* 27.10, pp. 3991–3994 *Cited on pages 45, 46.*

Xu, Yi, Licheng Yu, Hongteng Xu, Hao Zhang, and Truong Nguyen (2015). "Vector sparse representation of color image using quaternion matrix analysis". In: *IEEE Trans. on Image Processing* 24.4, pp. 1315–1329 *Cited on pages 12, 17, 46, 70.*

Yanase, Juri and Evangelos Triantaphyllou (2019). "A systematic survey of computer-aided diagnosis in medicine: Past and present developments". In: *Expert Systems with Applications* 138:112821 *Cited on page 2.*

Yang, Meng, Lei Zhang, Xiangchu Feng, and David Zhang (2011). "Fisher discrimination dictionary learning for sparse representation". In: *IEEE Int. Conf. on Computer Vision (ICCV), 2011*. IEEE. November 6-13, Barcelona, Spain, pp. 543–550 *Cited on page 82.*

— (2014). "Sparse representation based fisher discrimination dictionary learning for image classification". In: *Inter. J. of Computer Vision* 109.3, pp. 209–232 *Cited on pages 6, 12, 26, 28, 35, 36, 42, 44, 45, 62, 65.*

Yang, Meng, Lei Zhang, Jian Yang, and David Zhang (2010). "Metaface learning for sparse representation based face recognition". In: *17th IEEE Int. Conf. on Image Processing (ICIP), 2010*. IEEE. September 26-29, Hong Kong, pp. 1601–1604 *Cited on page 6.*

Yao, Tingting, Zhiyong Wang, Zhao Xie, Jun Gao, and David Dagan Feng (2016). "A multiview joint sparse representation with discriminative dictionary for melanoma detection". In: *2016 Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. November, Gold Coast, Australia, pp. 1–6 *Cited on page 91.*

Yin, Ming, Wei Liu, Jun Shui, and Jiangmin Wu (2012). "Quaternion wavelet analysis and application in image denoising". In: *Mathematical Problems in Eng.* 2012 *Cited on pages 45, 51.*

Yu, Zhen, Xudong Jiang, Feng Zhou, Jing Qin, Dong Ni, Siping Chen, Baiying Lei, and Tianfu Wang (2018). "Melanoma recognition in Dermoscopy images via aggregated deep convolutional features". In: *IEEE Trans. on Biomedical Engineering* 66.4, pp. 1006–1016 *Cited on page 100.*

Zagoruyko, Sergey and Nikos Komodakis (2016). "Wide residual networks". In: *arXiv preprint arXiv:1605.07146* *Cited on pages 75, 76.*

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European Conf. on Computer Vision*. Springer. September, Zurich, Switzerland, pp. 818–833 *Cited on pages 20, 22.*

Zhang, Haichao, Yanning Zhang, Nasser M Nasrabadi, and Thomas S Huang (2012). "Joint-structured-sparsity-based classification for multiple-measurement transient acoustic signals". In: *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.6, pp. 1586–1598 *Cited on page 6.*

Zhang, Heng, Vishal M Patel, Sumit Shekhar, and Rama Chellappa (2015). "Domain adaptive sparse representation-based classification". In: *2015 11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. May, Ljubljana, Slovenia, pp. 1–8 *Cited on pages 9, 82, 83, 85.*

Zhang, Jianpeng, Yutong Xie, Yong Xia, and Chunhua Shen (2019). "Attention residual learning for skin lesion classification". In: *IEEE Trans. on Medical Imaging* 38.9, pp. 2092–2103 *Cited on pages 91–93, 96.*

Zhang, Qiang and Baoxin Li (2010). "Discriminative K-SVD for dictionary learning in face recognition". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010*. IEEE. June 13-18, San Francisco, CA, USA, pp. 2691–2698 *Cited on pages 6, 12, 26–28, 36, 44, 45.*

Zhang, Shiqing, Lemin Li, and Zhijin Zhao (2012). "Facial expression recognition based on Gabor wavelets and sparse representation". In: *IEEE 11th Int. Conf. on Signal Processing*. Vol. 2. IEEE. October, Beijing, China, pp. 816–819 *Cited on page 29.*

Zhang, Zheng, Yong Xu, Jian Yang, Xuelong Li, and David Zhang (2015). "A survey of sparse representation: algorithms and applications". In: *IEEE access* 3, pp. 490–530 *Cited on pages 4, 27.*

Zhou, Zhuhuang, Shuicai Wu, King-Jen Chang, Wei-Ren Chen, Yung-Sheng Chen, Wen-Hung Kuo, Chung-Chih Lin, and Po-Hsiang Tsui (2015). "Classification of benign and malignant breast tumors in ultrasound images with posterior acoustic shadowing using half-contour features". In: *J. of Medical and Biological Engineering* 35.2, pp. 178–187 *Cited on page 2.*

Zhu, Wenwu, Xin Wang, and Wen Gao (2020). "Multimedia intelligence: When multimedia meets artificial intelligence". In: *IEEE Trans. on Multimedia* 22.7, pp. 1823–1835 *Cited on page 2.*

Zou, Cuiming, Kit Ian Kou, and Yulong Wang (2016). "Quaternion collaborative and sparse representation with application to color face recognition". In: *IEEE Trans. on Image Processing* 25.7, pp. 3287–3302 *Cited on pages 12, 17, 18, 46, 56–59, 63, 70, 92.*

Zou, Weibao and Yan Li (2007). "Image classification using wavelet coefficients in low-pass bands". In: *Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE. August, Orlando, Florida, USA, pp. 114–118 *Cited on pages 29, 30, 45, 54, 76, 100.*

Zou, Weibao, King Chuen Lo, and Zheru Chi (2006). "Structured-based neural network classification of images using wavelet coefficients". In: *Int. Symposium on Neural Networks*. Springer. May, Chengdu, China, pp. 331–336 *Cited on page 4.*