Dictionary Learning for Pattern Classification in Medical Imaging

Hrishikesh Deshpande

ANNÉE 2016

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale MATISSE

présentée par

Hrishikesh DESHPANDE

Préparée à l'unité de recherche IRISA UMR CNRS 6074 / INRIA Rennes
Nom développé de l'unité : VisAGes - INSERM U746
UFR Informatique et Electronique (ISTIC)

Dictionary Learning for Pattern Classification in Medical Imaging

Thèse soutenue à Rennes
le 8 Juillet 2016

devant le jury composé de :

Daniel RUECKERT
Professor, Imperial College London / *Rapporteur*

Carole LARTIZIEN
Chargée de recherche, CNRS / *Rapporteur*

Alexandre GRAMFORT
Maître de Conférences, CNRS LTCI/ *Examinateur*

Rémi GRIBONVAL
Directeur de Recherche, INRIA / *Examinateur*

Christian BARILLOT
Directeur de Recherche, CNRS / *Directeur de thèse*

Pierre MAUREL
Maître de Conférences, Université de Rennes 1 /
Co-directeur de thèse

Acknowledgments

First and foremost, thanks to the God, the Almighty, for his blessings throughout this research work.

I would like to thank my thesis advisors Dr. Christian Barillot and Dr. Pierre Maurel for putting their trust in me and inviting me to do a Ph.D. in France. They guided me all along this work providing me with a certain amount of independence. The fruitful conversations we had during this period have helped me develop the thematic outline for this thesis. I am very grateful for their motivation, guidance and patience. I would also like to thank them for providing me an opportunity to visit Duke University, during this work.

Thanks to Prof. Guillermo Sapiro and Dr. Qiang Qiu for their scientific advice and knowledge, and many insightful discussions during my visit to the Duke University. They were a great source of knowledge and I hope that I could be as lively and energetic as them.

I am thankful to the jury members of my thesis committee. I thank Prof. Daniel Rueckert and Dr. Carole Lartizien, for reviewing this Ph.D. work and providing me with useful comments and suggestions. I also thank Dr. Rémi Gribonval, Dr. Alexandre Gramfort and my Ph.D. supervisors - Dr. Christian Barillot and Dr. Pierre Maurel, for participating to the committee of my PhD defense. I thank them all for their valuable time and suggestions.

I would like to thank INRIA Rennes for providing full financial support for this work. The doctoral school MATISSE gave an additional financial support for my visit to Prof. Guillermo Sapiro's lab at Duke University, USA.

I extend my sincere thanks to Angelique Jarnoux, Emilie Gesnys and Mary Pope for their administrative support. I thank all my friends, the present and the past members of the team VISAGES at INRIA Rennes and the department of electrical and computer engineering at Duke University, USA.

Thanks to Prof. V. M. Gadre at Indian Institute of Technology Bombay, Dr. S. T. Hamde, Dr. B. M. Patre and Dr. R. S. Holambe at Shri Guru Gobind Singhji College of Engineering and Technology Nanded, and my seniors Dr. Bhushan Patil and Dr. Sajan Goud, who motivated me throughout my education and played a big role in shaping my career.

Finally, I would like to thank my family for their unconditional support and sacrifices. My mother (Mrs. Chhaya Deshpande) and father (Mrs. Narayan Deshpande) have always been my source of inspiration and they always encouraged me for better education. I thank my wife (Mrs. Mukti Sadhu) for her constant support, encouragement and understanding. To my sister (Miss. Rutuja Deshpande), I am grateful for bringing me so much joy and love. I thank my father and mother-in-laws and all my relatives who indirectly contributed in my well-being and this work.

Contents

Chapter 1

Résumé en français

Depuis plusieurs décennies, la quantité de données générées et stockées est en hausse exponentielle. Les téléphones portables collectent des données telles que des images, des paroles, des battements de coeur, nombre de pas, ... Les satellites capturent des données relatives aux informations météo. Des millions d'utilisateurs téléchargent d'énormes quantités d'informations sur les réseaux sociaux tels que Facebook, Twitter, ... Les dispositifs médicaux acquièrent des images haute résolution du corps humain. Ces données peuvent être utiles aux organismes collecteurs et trouver les contenus utiles dans les données est une étape essentielle dans la prise de décision future. La visualisation de ces données par l'humain afin de trouver les motifs pertinents est rendue difficile par la grande dimensionnalité des données. D'autre part, l'analyse des interactions entre un grand nombre de variables va au-delà des capacités des experts humains.

Afin d'améliorer la compréhension de ces problèmes, le domaine de l'apprentissage automatique a beaucoup évolué. Il a profité de l'augmentation des capacités de calcul des machines afin de découvrir des motifs cachées et de faire des prédictions sur les données sans programmer de manière explicite les algorithmes. Les algorithmes d'apprentissage automatique sont capables de fournir des solutions pour les problèmes de grande dimension avec une bonne reproductibilité. Cette connaissance peut être transférée et mise à l'échelle à travers de multiples applications et pour des millions d'utilisateurs sans intervention humaine. Quelques applications notables de l'apprentissage automatique et de la reconnaissance des formes sont le filtrage de spam, les systèmes de navigation et de guidage, les moteurs de recherche, la vision par ordinateur et des systèmes de recommandation tels que Netflix, Amazon, etc. Dans le système de recommandation, par exemple, une liste de recommandations de films ou de produits est suggérée à un utilisateur en utilisant le modèle appris du comportement passé de l'utilisateur ou des décisions similaires faites par d'autres utilisateurs.

Plusieurs méthodes d'apprentissage automatique ont été proposées au cours des dernières années, qui nécessitent des données d'apprentissage étiquetées, ou qui explorent les données non étiquetées pour trouver des structures en leur

sein. Le succès de l'algorithme d'apprentissage choisi dépend en grande partie des fonctions utilisées ainsi que de la distribution statistique sous-jacente. Par exemple, les modèles de mélanges gaussiens supposent que les données observées se composent d'un mélange de plusieurs gaussiennes. Tout déviation par rapport à cette hypothèse pourrait entraîner des performances de classification détériorées. De même, l'utilisation d'une fonction de classification linéaire dans le cas où il existe une interaction non linéaire entre les prédicteurs pourrait conduire à des résultats de classification dégradés. La modélisation des données joue donc un rôle important dans le choix des algorithmes d'apprentissage automatique et la qualité de la classification.

Récemment, la modélisation du signal en utilisant des représentations parcimonieuses a suscité un intérêt croissant. Les signaux naturels et les images peuvent être représentés par une combinaison linéaire d'un petit nombre de coefficients en utilisant une famille de fonctions de base organisées dans les colonnes d'un dictionnaire. L'utilisation d'un dictionnaire fixe tels que les ondelettes permet un calcul rapide des coefficients parcimonieux mais un tel dictionnaire offre une capacité d'adaptation limitée. Avec l'avènement des méthodes d'apprentissage automatique, il est devenu possible d'apprendre un dictionnaire adapté aux données, améliorant ainsi la capacité d'adaptation de données. Ces dictionnaires sont connus pour avoir un meilleur pouvoir de représentation et leur utilisation a permis d'améliorer les performances d'applications telles que le débruitage d'images, la restauration, l'inpainting, etc. Au cours des dernières années, l'apport de ces méthodes d'apprentissage de dictionnaires dans la classification d'images a été étudié. Ces approches étendent souvent le cadre de l'apprentissage de dictionnaires standard, de sorte que les dictionnaires soient discriminatifs en plus d'être représentatifs. Ces algorithmes sont utilisés avec succès dans des applications telles que la classification d'images, la catégorisation, la segmentation etc. La recherche dans la communauté parcimonieuse a été axée sur trois aspects différents: (i) le développement de méthodes efficaces pour le calcul de représentations parcimonieuses, (ii) l'élaboration de méthodes d'apprentissage de dictionnaires pour la représentation et la classification, et (iii) l'exploration de l'utilisation d'apprentissages de dictionnaires dans diverses applications.

Dans cette thèse, nous étudions le rôle des représentations parcimonieuses et d'apprentissage de dictionnaires en reconnaissance de motif. Tout d'abord, nous proposons une méthode de classification, basée sur l'apprentissage de dictionnaires, qui prend en considération les différences de complexité entre les différentes classes. Les motifs d'intérêt à classer sont souvent moins fréquents et sont associés à une faible variabilité par rapport à la structure de fond. Apprendre des dictionnaires spécifiques à chaque classe aboutit à une bonne

puissance de représentation, mais ne garantit pas une bonne classification. Les informations de variabilité entre les classes pourraient être utilisées efficacement pour ajouter de la puissance de discrimination aux dictionnaires. Nous validons notre approche sur une application de vision par ordinateur, la détection des lèvres dans des images de visage.

Nous proposons également une application de l'apprentissage de dictionnaires à la classification d'images médicales. Bien que les techniques de représentations parcimonieuses et d'apprentissage de dictionnaires soient largement utilisés en vision par ordinateur (reconnaissance faciale, classification de textures, reconnaissance d'actions) leur utilisation dans le domaine de l'imagerie médicale n'a commencé à croître que récemment. Nous abordons un problème cliniquement pertinent: détecter des motifs pathologiques, des lésions de scléroses en plaques (SEP), dans des images multimodales IRM (imagerie par résonance magnétique) de cerveaux. La délimitation manuelle des lésions de SEP nécessite des experts en neuro-radiologie et l'analyse multimodales d'images IRM est une tâche laborieuse et prend du temps. En outre, la grande hétérogénéité, en forme et en intensité, des lésions de SEP entraîne des différences de segmentation intra- et inter-experts. Notre approche aborde la question de traiter un grand volume d'images IRM multimodales, de façon automatisée, et réalise la classification de lésions SEP en tenant compte des différences de complexité entre les motifs pathologiques à identifier (lésions SEP) et les structures cérébrales saines telles que la matière blanche, la matière grise et le liquide céphalo-rachidien, en arrière-plan.

Organisation de la thèse

Cette thèse est organisée en deux parties. La première partie se compose de trois chapitres qui présentent le contexte et la motivation de notre travail, ainsi qu'une introduction à l'apprentissage automatique, la reconnaissance des formes et aux représentations parcimonieuses. En particulier, le chapitre 2 présente l'architecture et l'organisation de la thèse. Le chapitre 3 décrit les concepts de base dans l'apprentissage automatique et la reconnaissance des formes et présente quelques algorithmes populaires ainsi que quelques applications. Le chapitre 4 porte sur la modélisation parcimonieuse d'images ainsi que sur l'apprentissage de dictionnaires. L'idée générale du cadre parcimonieux et de l'apprentissage de dictionnaires est exposée et quelques algorithmes populaires sont décrits. Nous avons utilisé quelques-uns de ces algorithmes pour la comparaison avec les méthodes présentées dans la prochaine partie de la thèse.

Dans la deuxième partie, trois chapitres couvrent les contributions faites

dans cette thèse, les expériences réalisées et les résultats. Le chapitre 5 expose la motivation derrière notre travail, pourquoi la taille des dictionnaires utilisés pour la classification pourrait jouer un rôle important. Nous démontrons l'importance de la taille de dictionnaire dans une application de vision par ordinateur, la détection des lèvres dans des images de visage, où il y a d'énormes différences de variabilité entre chaque classe. Dans le chapitre 6, nous étudions l'utilisation de représentations parcimonieuses et l'apprentissage de dictionnaires dans une application plus complexe, concernant l'imagerie médicale. Nous abordons le problème cliniquement pertinent de la classification d'une pathologie du cerveau (la SEP) à l'aide d'images IRM multimodales. Enfin, le chapitre 7 conclut la thèse par des perspectives sur le travail accompli.

Contributions

Chapitre 3: Apprentissage automatique et reconnaissance de formes

L'apprentissage automatique est un des domaines les plus actifs de l'informatique et a joué un rôle crucial dans des domaines aussi variés que l'automatisation, la médecine, les finances, etc. Plusieurs algorithmes d'apprentissage automatique ont été proposés au cours des dernières décennies. Ils peuvent être classés en deux types: (i) l'apprentissage supervisé: Les données sont présentées à l'ordinateur avec des exemples de couple (entrées, sorties), et l'objectif est d'apprendre une règle qui fait correspondre les entrées aux sorties, et (ii) l'apprentissage non supervisé : il n'y a pas d'étiquettes disponibles pour l'ensemble des données fournies et l'objectif est de trouver les motifs cachés ou des structures dans les données. Dans ce chapitre, nous présentons quelques principes fondamentaux dans l'apprentissage automatique, fournissons quelques exemples où cette technologie est utilisée et discutons des critères de performance pour le développement d'algorithmes. Dans ce chapitre, nous présentons également les concepts de base en reconnaissance des formes, une branche de l'apprentissage automatique qui met l'accent sur la reconnaissance des motifs et des régularités à partir d'un ensemble de signaux numériques ou des images. Nous rencontrons des exemples dans la vie de tous les jours, comme la reconnaissance d'empreintes digitales, la reconnaissance vocale dans les téléphones portables, etc.

Chapitre 4: Représentations parcimonieuses et apprentissage de dictionnaires

Les représentations parcimonieuses permettent aux signaux d'être représentés par une combinaison linéaire de quelques atomes dans un dictionnaire de plus grande dimension. La représentation du signal de cette manière a sus-

cité un vif intérêt au cours des dernières années car la plupart des signaux naturels et des images admettent des représentations parcimonieuses dans des bases fixes telles que Fourier, les ondelettes, etc. Les dictionnaires appris à partir des données se sont révélés être plus efficace que les dictionnaires fixes, qui ont une capacité d'adaptation limitée en raison d'une formulation mathématique explicite. Ces méthodes qui apprennent des fonctions de base non-paramétriques et qui donnent lieu à une représentation parcimonieuse des données sont appelées méthodes d'apprentissage de dictionnaires. Nous discutons quelques approches notables proposées au cours des dernières années pour trouver les coefficients parcimonieux et pour apprendre les dictionnaires plus adaptés à application donnée, comme le débruitage d'images, l'inpainting, etc. La partie suivante du chapitre décrit les différentes méthodes mises au point pour la classification d'images à l'aide d'apprentissage de dictionnaires et des techniques de représentation parcimonieuses. L'objectif principal de ces approches est d'apprendre des dictionnaires qui conduisent à une meilleure représentation des données, mais aussi à une meilleure discrimination entre classes. Enfin, quelques applications notables des techniques d'apprentissage de dictionnaires, tels que le débruitage d'images, la compression, la classification, etc., sont exposés.

Chapitre 5: Rôle de la taille des dictionnaires pour la classification

Il existe plusieurs méthodes de classification des images utilisant l'apprentissage de dictionnaires, mais celles-ci présentent plusieurs inconvénients, à la fois en ce qui concerne l'apprentissage classique de dictionnaire et l'apprentissage de dictionnaires discriminatifs. D'une part, les approches d'apprentissage de dictionnaires classiques utilisent des dictionnaires spécifiques à chaque classe, mais qui ne tiennent pas compte de la variabilité inter-classes. Les approches d'apprentissage de dictionnaires discriminatifs, d'autre part, nécessitent généralement des calculs excessivement lourds et présentent un grand nombre de paramètres qui doivent être ajustés pour le problème considéré. Nous proposons une méthode de classification qui tient compte des différences de variabilité entre les motifs à classifier et les informations d'arrière-plan, en utilisant des dictionnaires de tailles différentes pour chaque classe. Nous discutons d'abord pourquoi la taille des dictionnaires est cruciale dans les applications de reconnaissance des formes où il y a grande variabilité entre les données des différentes classes et démontrons en outre l'importance de la taille des dictionnaires dans une application en vision par ordinateur particulière: la détection des lèvres dans des images de visage. Une information a priori de différences de variabilité entre la classe "lèvres" et la classe "non-lèvres" est utilisée efficacement dans le cadre de l'apprentissage des dictionnaires, en

incorporant différentes tailles de dictionnaire pour chaque classe. Nous insistons sur le fait que la taille du dictionnaire n'est pas simplement un paramètre parmi d'autres, mais il commande directement deux propriétés fondamentales des dictionnaires utilisés dans la classification: la puissance de représentation des données et la capacité de discrimination inter-classes. Le choix de la taille des dictionnaires est une question clé dans l'amélioration de la classification d'images. Nous étudions la sélection des tailles de dictionnaire pour obtenir une classification optimale en utilisant trois approches différentes: (i) l'Analyse par Composantes Principales (ACP): les différences de complexité des données entre classes sont étudiées en utilisant le nombre de vecteurs propres nécessaires pour atteindre une valeur particulière de variance cumulée pour chaque classe. (ii) Des mesures basés sur les histogrammes d'erreurs : les dictionnaires appris pour chaque classe sont analysés pour obtenir les histogrammes des erreurs de reconstruction et la taille optimale de chaque dictionnaire est sélectionnée lorsque le même niveau de représentativité est atteint pour chaque classe, et (iii) la sélection empirique des tailles de dictionnaires pour chaque classe permettant d'atteindre le meilleur taux de classification sur l'ensemble d'apprentissage.

Chapter 6: Classification of Multiple Sclerosis Lesions

La sclérose en plaques (SEP) est une maladie démyélinisante auto-immune du système nerveux central et est l'une des principales causes d'handicaps physiques et cognitifs chez les jeunes adultes. L'IRM s'est révélé être la meilleure technique d'imagerie pour le diagnostic des lésions de sclérose en plaques dans le cerveau et est largement utilisé en clinique pour l'observation, le pronostic de la maladie et l'efficacité du traitement. L'analyse visuelle d'un grand nombre d'images IRM multimodale permet de mettre en évidence les lésions de SEP, mais est une tâche fastidieuse et sujette à une grande variabilité inter- et intra-experts. Dans ce chapitre, nous commençons par lister les approches de segmentation automatiques de lésion SEP proposées au cours des dernières années et les classons en techniques supervisées ou non supervisées. Dans la partie suivante, nous proposons une approche supervisée pour la classification des lésions SEP. Ceci est réalisé par l'apprentissage de dictionnaires spécifiques au le tissu cérébral sain et aux lésions, et en permettant différentes tailles de dictionnaire pour chaque classe, afin de prendre en compte les différences de variabilité entre les lésions SEP et les tissus cérébraux sains plus complexes. Nous étudions de nouveau le problème du choix de la taille des différents dictionnaires à l'aide de l'ACP et des mesures basées sur les histogrammes des erreurs. On observe que l'ACP n'est pas capable de fournir précisément le rapport entre la taille de chaque dictionnaire pour les deux classes, probablement en raison des structures non linéaires présentes dans

les données de la classe "tissus sains". Ce problème est résolu par la subdivision de cette classe pour chaque tissu cérébral sain, substance blanche, matière grise et liquide céphalo-rachidien, au lieu d'apprendre un seul dictionnaire pour la classe combinée. Les distributions gaussiennes sous-jacentes de chaque tissu cérébral sain permettent à l'ACP de fournir les tailles de dictionnaires optimales. Enfin, le rôle de la taille des dictionnaires dans l'une des approches les plus populaires d'apprentissage de dictionnaires discriminatifs, Fisher Discrimination Dictionary Learning (FDDL), a été étudié dans la classification des lésions de SEP.

Chapitre 7: Conclusion

Dans cette thèse, nous avons étudié le rôle des représentations parcimonieuses et de l'apprentissage de dictionnaires dans les applications de classification de formes, où il existe des différences de variabilité entre classes. Nous avons découvert qu'une amélioration majeure dans la classification de motifs peut être obtenue en adaptant la taille des dictionnaires pour chaque classe, à la fois dans le cas des dictionnaires classiques et des dictionnaires discriminatifs. Nous affirmons que la taille des dictionnaires n'est pas simplement un paramètre parmi d'autres, en particulier à des fins de classification où l'on compare la puissance de représentation de plusieurs dictionnaires. Pour illustrer le caractère générique de cette affirmation, nous avons validé la proposition d'utiliser différentes tailles de dictionnaires dans une application de vision par ordinateur, la détection des lèvres dans des images de visages, ainsi que par une application médicale plus complexe, la classification des lésions de scléroses en plaques dans des images IRM multimodales.

Chapter 2
Introduction

Since the last few decades, the amount of data being generated and stored is rising exponentially. The mobile sensors collect data such as pictures, audio signals, biological parameters such as heart rate etc., the satellites revolving around the earth capture the data pertaining to the weather information, millions of users upload huge amount of information on social networking sites such as Facebook, Twitter etc., medical acquisition devices obtain high resolution images of a human body. This data can be valuable to the organizations collecting it and finding the useful contents in the data is a vital step in further decision making. Visualization of such data by humans in order to find the relevant patterns is made difficult by high dimensionality of the data. On the other hand, analyzing interactions between large number of variables goes beyond the capabilities of human experts.

To improve the understanding of such problems, the field of machine learning has evolved from the study of pattern recognition and artificial intelligence. It takes advantage of increased computational capabilities of machines in order to find the hidden insights and make predictions on data without explicitly programming the computers. Machine learning algorithms are capable of providing solutions for high-dimensional problems with good reproducibility. This knowledge can be transferred and scaled across multiple applications and millions of users without any or minimal need of human intervention. Few notable applications of machine learning and pattern recognition include spam filtering, navigation and guidance systems, search engines, computer vision and recommender systems such as Netflix, Amazon etc. In recommender system, for example, a list of recommendations of movies or products is suggested to a user with the help of model learned from the past behavior of the same user or similar decisions made by other users.

Several machine learning techniques have been proposed over the past few years, which either require a labelled training data or explore unlabelled data to find some structures within the given data set. The success of the selected machine learning algorithm largely depends on the features used as well as distribution of the underlying data. For example, a popular machine learning approach known as Gaussian mixture model assumes that the observed data is composed of a mixture of several Gaussian distributions. Any deviation

from this assumption might result in deteriorated classification performance in the given application. Similarly, the use of a linear classification function in the cases where there is non-linear interaction among predictors could lead to worse classification results. The data modelling thus plays an important role in the choice of machine learning algorithms and the classification accuracy.

Recently, the signal modelling using sparse representations has gained a special attention. The natural signals and images can be represented by a linear combination of few coefficients using a set of basis functions organized as the columns of a dictionary. The use of a fixed dictionary such as Wavelets results in faster computation of sparse coefficients but such dictionary offers limited adaptability on account of fixed mathematical formulation in employing these basis functions. With the advent of machine learning methods, it became possible to learn the dictionary from the underlying data so that the best set of basis functions could be learned for obtaining the sparse representation of the data, thus improving the data adaptability. Such dictionaries have a good representation power and their use has resulted in improved performance in image processing applications such as denoising, restoration, inpainting etc, instead of using fixed dictionaries. In the last few years, several researchers have investigated the use of dictionary learning technique in image classification. These approaches often extend the standard dictionary learning framework so that the dictionaries are discriminative in addition to being representative of their class data. Such algorithms are successfully used in developing applications such as image categorization, segmentation etc. The research in sparsity community has been focused on three different aspects: (i) the development of efficient methods for calculating sparse representations, (ii) proposition of dictionary learning algorithms for signal representation and classification, and (iii) investigate the use of dictionary learning and sparse representation paradigm in various applications.

In this thesis, we investigate the role of sparse representations and dictionary learning technique in pattern recognition applications. Firstly, we propose the dictionary learning based classification approach which takes into consideration the complexity differences between class data. The patterns of interest are less occurring phenomenon and are less complex structures as compared to the background information. Learning class specific dictionaries results in good representation power, but it does not guarantee best classification. The variability information between class data could be effectively used to add discrimination power into the dictionaries. We validate our approach using a computer vision application such as lips detection in face images.

In the next part, we propose an application of dictionary learning and sparse representation based classification method in medical imaging. While

sparse representation and dictionary learning techniques are widely used in computer vision applications such as face recognition, texture classification and activity recognition etc, their use in the field of medical imaging has started growing only recently. We address a clinically relevant problem of classifying pathological patterns called Multiple Sclerosis (MS) lesions in multi-channel brain Magnetic Resonance (MR) images using the sparse representation and dictionary learning technique. The manual delineation of MS lesions requires neuro-radiological experts and analyzing multi-channel MR images is a laborious and time consuming task. Furthermore, huge heterogeneity in the shape and intensity patterns of MS lesions leads to intra- and inter-rater segmentation differences. Our approach addresses the issue of processing a huge volume of multi-channel MR images and achieves the MS lesions classification by considering variability differences between the patterns to be identified (MS lesions) and the background brain structures (White matter, grey matter and cerebrospinal fluid).

This thesis is organized as follows:

Chapter 3: Machine Learning and Pattern Recognition

Machine learning has played a crucial role in the fields as diverse as automation, medicines, finance etc. The field of pattern recognition is employed in automatic detection of patterns from a set of digital signals or images and we come across its examples in day-to-day life, such as fingerprint recognition, speech recognition etc. In this chapter, we introduce some fundamentals in machine learning and pattern recognition. We provide some examples where this technology is used and discuss the performance criteria for the development of algorithms using this technology.

Chapter 4: Sparse Representations and Dictionary Learning

Sparse representation allows the signals to be represented by a linear combination of few atoms in an over-complete dictionary. We discuss few notable approaches proposed over the last few years for finding the sparse coefficients and to learn the dictionaries better suited for a given application such as image denoising, inpainting etc. The last part of the chapter describes various methods developed for obtaining image classification using advanced dictionary learning techniques, known as discriminative dictionary learning.

Chapter 5: Role of Dictionary Size in Pattern Classification

There exist several methods for image classification using dictionary learning, but they are associated with several disadvantages in the case of both the standard and discriminative dictionary learning techniques. On one hand, the

standard dictionary learning approaches use class specific dictionaries, which does not take into account the variability between class data. The discriminative dictionary learning approaches, on the other hand, are computationally demanding and are associated with a large number of parameters which need to be tuned for the given pattern recognition problem. We propose a classification method which takes into consideration the variability differences between the patterns to be classified and the background information by employing the dictionaries of different size for each class. Finally, we demonstrate the significance of dictionary size in a particular computer vision application such as lips detection in face images, in the case of both the standard and the discriminative dictionary learning methods.

Chapter 6: Classification of Multiple Sclerosis Lesions

Multiple Sclerosis is an autoimmune, demyelinating disease of the central nervous system and is one of the main causes for developing physical and cognitive disabilities in young adults both in developed and developing world. MRI has proved to be the best paraclinical imaging technique for the diagnosis of MS lesions in the brain and is widely used in the clinical setting for observing the disease prognosis and the treatment efficiency. We proposed the dictionary learning based MS lesions classification technique by using the class specific dictionaries of different sizes for the healthy brain tissues and the MS lesions class. Finally, an adaptive dictionary learning method is proposed by learning the dictionaries for each healthy brain tissue - White matter, grey matter and cerebrospinal fluid, and the lesions class, while principal component analysis of the data and histogram based measures from the learned dictionaries are used to select the size of the dictionary for each class.

Chapter 7: Conclusion

This chapter summarizes the perspectives on the problems addressed in the thesis and provides the conclusions and contributions.

Chapter 3

Machine Learning and Pattern Recognition

Contents

Machine learning is a multidisciplinary field, which has mainly emerged from artificial intelligence, computer science and applied mathematics. Over the past few decades, machine learning has received a great deal of attention and today, there exist numerous successful applications using this technology. The field has not only grown in terms of significant theoretical contributions but has also found practical applications in the fields as diverse as finance, biology, medicine, robotics, arts, entertainment etc. With the generation of more and more digital data, advancements in the computational power of the machines and rapidly growing community, the filed of machine learning has made a transition from the laboratory demonstrations to critical real-world applications and has attained a significant commercial value.

3.1 What is Machine Learning?

The goal of machine learning is to make computers able to *learn*. Machine learning, in general, refers to learning patterns and characteristic structures from the data in order to make predictions and decisions on unseen data of similar type. This enables computers to take decisions based on the provided data, instead of explicitly programming them to carry out a dedicated

task. The machine learning algorithms also have capability to learn and improve over time when exposed to new data. Machine learning algorithms are designed to infer unknown variable values corresponding to the given independent variable(s) and the data.

A few notable applications of machine learning are recommender systems, spam detection, web page ranking, face recognition, natural language processing, climate modeling, sentiment analysis, medical diagnosis etc. We illustrate few such examples below:

- Consider collaborative filtering used in the online shopping application such as Amazon, with an objective to obtain a sorted list of the product recommendations to a particular user, based on the purchase history and product views of the user. The decisions made by the similar users (hence the term 'collaborative') can be used to *learn* how to predict the future purchase or recommend products for viewing to the user. The machine learning approach in this particular application provides a clear advantage in handling a huge number of users in recommending products, which is impossible to be done manually.

- E-mail spam detection is a classification problem for deciding whether an e-mail contains relevant information or not. This is a user-dependent problem: Frequent e-mails from a particular service notifying discounts might be a valuable information for one user but it might not be a similar case for other users. Thus, the classification method should consider user preferences and it should have a capability to adapt over time, as the preferences of each user might change over time. We can process the contents of e-mail to generate word counts for each e-mail and design a binary classifier for spam detection using previous knowledge of frequently occurring word counts for spam mails and user preferences.

- For a critical application such as credit card fraud detection, the anomaly detection technique is used. This type of classification technique deals with detection of fraudulent transactions as outliers with respect to the normal purchasing pattern from the user.

3.1.1 Advantages and Disadvantages of Machine Learning

Machine learning offers several advantages as described below:

1. Enormous amount of data is being generated and stored. For example. twitter, online shopping, medical images etc. To automatically identify

and process the most relevant content in these huge data sets is one aim behind machine learning techniques. Smart data analysis will play an important role in the technological advancements in the years to come.

2. Analyzing such large data sets require human efforts which are prone to error. Minimizing human involvement in such cumbersome tasks, while maintaining a good accuracy in performing such tasks is another advantage of machine learning algorithms.

3. Humans are prone to error while dealing with multivariate data as it is complicated to find out the relationship among several features manually. Machine learning algorithms can be incorporated in such problems, for improving the accuracy and efficiency of the application.

4. A high-dimensional data, such as medical images, is often complicated to analyze for humans and this usually requires skilled persons. In addition, such complex tasks can be time-consuming. Machine learning algorithms provide good alternative to such repetitive tasks, saving human labor and time.

5. Changes in the user preferences or new knowledge in the training data might ask for redesign of the system. Machine learning techniques can adapt to these changes in a better manner.

6. In applications such as face recognition, it is impossible to define handwritten rules. Machine learning algorithms offer effective solutions in such scenarios.

7. With the advent of high-performance machines, it is now possible to distribute the data and process it in a manner that was not possible few years ago. Such improvement in machine performance allows storing and processing big data in the shortest possible time. GPU based implementations, along with parallel programming and distributed systems have allowed to develop more powerful machine learning methods in recent years.

8. Neurologists or radiologists are relatively scarce in numbers as compared to the population of the patients they cater to. Furthermore, rural population in some countries has a little access to such highly-skilled experts. Machine learning techniques, coupled with other technological advancements can play a big role in bridging the gap between the patients and the doctors.

However, there are some disadvantages associated with machine learning techniques, as mentioned below:

1. Supervised machine learning approaches need a lot of labelled data. For example, in the case of sentiment analysis, to predict whether tweets are associated with a positive, negative or neutral sentiment, one needs to label each tweet in training data set with either of the target classes. This requires human efforts and is a time-consuming task. In addition, the task of labelling the training data can get complicated in applications such as medical imaging, where skilled personnel are needed.

2. Machine learning algorithms are not guaranteed to always work in every case imaginable. The domain knowledge of the problem at hand is necessary to apply the right machine learning algorithm.

3. The use of machine learning might raise ethical issues when important decisions are taken from a machine learning algorithm and is applied to a wrong individual. For example, a health insurance company detects a risk from a large population and associates the risk to you even if you are a false positive. An ethical issue in this case is who should be hold responsible for wrong decision?

3.1.2 Machine Learning Approaches

Machine learning addresses several problems. For example, binary classification deals with separating a set of data points into two groups, based on selected features. Multiclass classification is an extension of binary classification, in which the instances are to be classified into more than two classes. A common approach to solve such problem is to convert a multiclass classification problem into multiple binary classification problems. The popular methods include one-vs-one and one-vs-all.

The machine learning approaches can be broadly classified into supervised and unsupervised learning.

3.1.2.1 Supervised Learning

In supervised learning, we are given with the data set $\{(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)\}$ and the goal is to estimate y for a given new value of $\mathbf{x}$. Here $\mathbf{x_i}$ represent the feature values or vectors (in the case of multivariate data), and y_i are the classes or target values. The objective of supervised learning is to infer a function, also called as a classifier, using given pairs of features and the desired output values, so that the inferred function predicts the correct class

for any valid unseen input. Based on whether y_i is a finite set or a continuous one, the problem can be viewed as classification or regression, respectively. In the classification problem, the target value belongs to either class and the objective is to predict the class labels for unseen data using all the information from the training data along with the labels of the classes. One example of this kind would be image categorization, where images are to be classified into different categories such as bird, airplane, house, river etc. On the other hand, the learning problem can be seen as regression when the target value is in the form of a continuous variable. Predicting housing prices using relevant house details which might include area, number of bedrooms etc. is an example of regression problem [Kotsiantis 2007].

In supervised learning approach, a model is learned so as to make predictions on the training data and is corrected when these predictions are wrong. The process is repeated until desired accuracy on training data is achieved. Examples of supervised learning algorithms include regression, neural networks and support vector machines, whereas the applications include spam detection, handwritten digit recognition, face recognition, sentiment analysis etc. [Burges 1998, Zhang 2000].

The input data, along with the label information, is known as training data and the ability of the classifier or regressor function to predict new unseen data is called generalization. Over-training of the model results in capturing every minute information in the training data and this might result in poor performance on unseen data, as the learned model is excessively tuned for the representation of the training data. Training error might keep decreasing with increment in the training iterations or the complexity of the model. Such model gives better performance on the training data, but the prediction error might start to deviate if too complex models are used. Such model is incapable of generalization and is said to overfit the training data. On the other hand, under-representation of training data leads to poor performance on the training as well as new unseen data, as incomplete information is captured with such low complexity models. It is therefore important in the case of supervised learning, to decide when to stop training the model from being too complex. An ideal supervised learning algorithm should be capable of learning complex functions and producing generalizable results.

3.1.2.2 Unsupervised Learning

In unsupervised learning, we are only given with the input vectors $\{(\mathbf{x_1}, ..., \mathbf{x_n})\}$, without any label information and the objective is to find the natural partitions or similar patterns in the underlying data. This technique is often used when key features or relations between the variables of the input data are to be

found. Most unsupervised learning techniques explore the idea of discovering similarities between vectors in the given data [Ghahramani 2004].

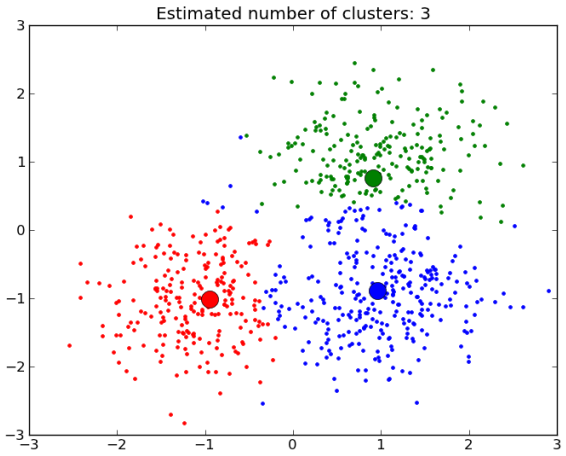


Figure 3.1: An example of clustering. The data points are divided into three clusters and each data point belonging to one of these clusters is shown in red, green or blue.

The most popular unsupervised learning algorithm is clustering, where the given set of patterns are divided into clusters in such a way that the data points belonging to same clusters exhibit similar properties and are part of the same class. One example of clustering is shown in Figure 3.1. Different algorithms exist to cluster the data. Hierarchical clustering creates a tree called dendrogram, which represents the data as a hierarchy of clusters. K-means clustering assigns each data point to one of K clusters in such a way that the sum of the euclidean distance between each data point and the centroid of its designated cluster is minimized. It is implemented as an iterative two-step procedure: The cluster assignment step assigns each data point to a cluster whose centroid is closer to the given data point, and the centroid update step calculates the new cluster center using arithmatic mean of all previously obtained assignments to the respective clusters. Gaussian Mixture Models (GMM) represent the given data as a mixture of multivariate normal distributions. These methods use a method called Expectation Maximization (EM), for the estimation of parameters of the models.

In density estimation, the data is assumed to belong to a particular probability distribution and the density or probability is found such that the member of a certain category will have particular features. This is difficult to achieve in higher dimensions. In such high-dimensional problems, another approach called dimensionality reduction is incorporated to find the lower dimensional

representation of the data, which approximately represents the given data. The fundamental assumption behind these techniques is that most of the information in a higher dimensional data lies on a lower dimensional manifold or union of manifolds, and this assumption is true for many real world applications. Examples of the dimensionality reduction techniques include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Locally Linear Embedding, Laplacian Eigenmaps etc.

3.1.2.3 Other Learning Methods

There are many other approaches which differ from supervised or unsupervised techniques discussed above.

In semi-supervised learning, the data set $\{(\mathbf{x_i}, y_1), ..., (\mathbf{x_k}, y_k), \mathbf{x_{k+1}}, ..., \mathbf{x_n}\}$ is a combination of labelled and unlabelled examples, and the target values for unlabelled variables is to be predicted. It is used in cases where labelling training data is expensive or scarse. Unlabelled data, on the other hand, is easier to collect, but there are few ways it could be used. Semi-supervised methods take the advantage of readily available unlabelled data to improve the supervised learning problem. Some popular semi-supervised algorithms include self-training, mixture models, multiview learning, graph-based methods, and semi-supervised support vector machines [Zhu 2005].

Another approach called active learning is a special case of semi-supervised learning in which an algorithm queries for the labels of particular points in the given data set. In such approach, the algorithm performs better with less training samples as the learner chooses the examples from which it learns [Settles 2010]. Reinforcement learning corresponds to designating rewards or losses corresponding to actions in the learning stages. The method performs learning by means of maximizing overall reward or minimizing loss.

Another interesting machine learning algorithms include ensembles of classifiers. Ensemble methods combine a set of classifiers and classify new data point by considering weighted or unweighted votes of the individual classifiers. The objective of this method is to improve the performance of individual classifiers and achieve improved generalizability. While using these classifiers, it is very important to analyze which base learners could be combined together and the methodologies to combine them. Different methods exist to create ensemble classifiers: Use different base classifiers, use different training parameters in a single base classifier or use different subsets of training data along with the same classifier. Most popular ensemble algorithms include boosting, bagging, random forests etc.

A set of input variables used as an input to the classifier is known as a feature vector. It is important to select the most representative features

that capture most of the information in the training data. Some machine learning applications require transformation of input data into a feature vector the classifier can understand. For example, in face recognition problem, the feature vector is obtained by transforming an image into vector, where each entry in the feature vector represents the intensity at each pixel in the face image. In some applications, it might be possible that only a subset of features are useful. Feature selection algorithms deal with the selection of most relevant features by scoring each feature.

3.1.3 Performance Metrics

Machine learning algorithms extract information from the provided training data, which might contain hundreds, thousands or millions of training samples, depending on application at hand. To test how well a machine learning algorithm performs, a subset of the given data set, called test data set, is prepared by selecting instances not contained in training data set. To evaluate the performance of the machine learning model, every data point in the test data set is given as an input to the model and the output of model is compared against the desired output for the corresponding input. The correctly identified test inputs are termed as True Positives (TP), incorrectly identified as False Positives (FP), correctly rejected as True Negatives (TN) and incorrectly rejected as False negatives (FN). The following measures are then used to evaluate the performance of machine learning algorithm.

Accuracy measures the proportion of correctly classified samples as a portion of total number of samples (L) given to the classifier.

$$Accuracy = \frac{TP + TN}{L} \tag{3.1}$$

Sensitivity (or recall) measures the proportion of positive samples that are classified correctly. Higher sensitivity indicates the ability of classifier to correctly detect test samples which actually belong to the positive class. However, sensitivity does not take into account false positive detections. Therefore, a machine learning algorithm predicting all given samples as belonging to positive class will have 100% sensitivity.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

Similarly, specificity measures the proportion of negative samples that are classified correctly. A classifier with a higher specificity indicates its ability to correctly detect test samples which actually belong to the negative class. A machine learning algorithm predicting no samples as belonging to positive class will have 100% specificity.

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

Positive Predictive Value (PPV) or precision is the proportion of correct positive classifications over all classifications assigned to the positive class.

$$PPV = \frac{TP}{TP + FP} \tag{3.4}$$

Similarly, Negative Predictive Value (NPV) can be calculated as follows.

$$NPV = \frac{TN}{TN + FN} \tag{3.5}$$

In addition to the above mentioned measures, there exist several other ways of indicating performance of the machine learning algorithm. The most popular among them is the confusion matrix, which indicates the actual and predicted number of test samples in tabular format.

3.2 Pattern Recognition

Humans are capable of sensing the surrounding environment, extract useful information and make the decisions based on this information. Humans, for example, undergo training in executing a particular task such as handwritten recognition and can classify newly seen digits even if there are slight variations in the newly observed data from the training data. This may include digits written on a variety of backgrounds of different color, texture or partially occluded digits. Pattern recognition is a scientific discipline in which machines observe the environment and automatically group or classify the measurements for making decisions or predictions. The research in last few decades has resulted in pattern recognition applications in the fields of artificial intelligence, communications, military intelligence, data mining, business, biology and medicine etc. Few notable application domains include computer vision, speech recognition, document classification, handwritten text analysis, medical diagnosis etc.

A pattern is a collection of measurements that are similar to one another in certain aspects. It describes the common trend within a set of measurements. A human face, ECG waveform, handwritten digit are all examples of patterns. Some complex applications might involve extracting statistical features from the underlying data. The patterns, in these cases, take the form of complex features. Individual patterns can be grouped together if they have similar properties. A good pattern recognition algorithm is one with as similar features as possible for a data belonging to a similar class and the most

discriminative features for the data between classes. This depends on how machines collect the information from the environment, identify the patterns of interest which are capable of distinguishing between different categories or classes and make decisions to classify the patterns.

Consider an example of Optical Character Recognition (OCR). The objective of this application is to automatically recognize the handwritten digits and characters in order to convert them into the text format. The algorithm should be able to assign each character or digit in the image to the corresponding set of output classes. The variations in the background, fonts and lighting complicate the recognition task, in addition to the different writing styles and slight rotations introduced in the input images. The pattern recognition algorithm should select the features and pre-processing steps, along with the design of a good classifier that is able to differentiate between the class data, while taking into account all the variations in the data set. Such a system might need specialized features extracted from the data, rather than simply using image intensities as the features. Furthermore, it is important to select the classifier that best captures the differences between class data and gives higher classification accuracy.

In the next subsections, we describe the general framework of pattern recognition systems, followed by few approaches proposed in the past.

3.2.1 General Framework

The design of pattern recognition system involves the following five steps.

3.2.1.1 Data Acquisition

The measurements of physical variables from the surrounding environment are collected with the help of sensors or digitizing machines. In some applications, the data is also acquired from the scanners. This step essentially converts the physical quantities into a form acceptable by computers for further processing. For example, the sound signal captured using microphone array, medical images acquired using MRI scanners, temperature data collected from thermal sensors etc.

This data might represent the interaction between many variables and the data set can be sub-categorized into different number of classes, depending on the application. For temperature data, the information received from sensors is representative of temperature variations with respect to time and can be categorized into classes like day and night temperatures or seasonal variations. Handwritten digit image data is an example of multi-variate data set, as this data arises from more than one variable.

3.2.1.2 Pre-processing

The data acquired is preprocessed for the removal of noise or isolating patterns of interest from the background. In some applications, the measurement data is segmented in such a way that each segmented object belongs to either class. For example in character recognition system, the image is searched for texts and numbers, which are then separated from background and individual characters, and these characters are then extracted for further processing. In patch-based segmentation methods, the images or volumes are subdivided into 2D or 3D blocks, which are assigned to a particular class based on some predefined rule. These individual elements can be represented as feature vectors.

3.2.1.3 Feature Extraction

This step involves the extraction of relevant features from the processed data. In some applications, such as face recognition, the dimensionality of a face image or a feature can be a large. A high definition face image can contain 512×512 pixels and the vector representation of this image, when used as a feature, leads to higher-dimensional input to the classifier. In addition, this might contain redundant information. One way to increase the efficiency of pattern recognition system is to reduce the number of features using dimensionality reduction technique such as Principal Component Analysis (PCA), while retaining as much input information as possible.

Some applications might require statistical or mathematical features such as mean, histogram or higher order statistics, while others can perform better when advanced image representation features such as wavelet sub-band energy are utilized as image features. There are other applications in which heuristically selected features help improve the classification. For example in classification of multiple sclerosis lesions, the area or shape information can prove to be useful in identifying lesions [Goldberg-Zimring 1998].

3.2.1.4 Classification

The classifier in the pattern recognition is a very important component, which receives feature vectors from the previous step as input, learns a model from the input data and assigns the new feature vector to the most appropriate class. The data set is often divided into two sets for this purpose. Training data set is used for learning a model, whereas test data is used for validating the efficiency of the system. The amount of data, types of features and the classifier, in principal, determine the efficiency of the pattern recognition system.

In text recognition system, the classifier receives the input images of an individual characters in the vector format and the output classes are recognized as one of the following classes: A, $B, ..., Z, 0, 1, ..., 9$ etc. The classifier is a mapping function from the input feature space to the set of classes. A good classifier has a better ability to distinguish feature vectors between classes.

Often, the classification and feature extraction steps are interlinked with each other. The features are extracted for better classification and the classifier tries to achieve the best classification with the given set of features. Each feature and classifier has several advantages and limitations associated with them and it is very important to understand which features and classifiers will give the best performance.

3.2.1.5 Post-processing

This optional step in pattern recognition refines the classification obtained in previous step to reduce the false detections and improve the performance by exploiting the context.

3.2.2 Methods for Pattern Recognition

Depending on the models used for classification, pattern recognition methods can be classified into following categories.

3.2.2.1 Statistical Pattern Recognition

In this approach, each pattern is represented as a feature vector of dimensionality d and these feature sets are chosen in such a way that patterns associated with different classes can be separated. The probability distributions of the patterns from each class of the training data are analyzed to determine the decision boundaries that separate the patterns from the different classes [Devroye 1996]. In supervised approach, discriminant analysis techniques like Linear Discriminant Analysis (LDA) or Fisher Discriminant Analysis (FDA) are used, where a discriminant function is defined which performs the classification. In the case of unsupervised learning, the approaches such as Principal Component Analysis (PCA) are used, in which, the patterns are detected in terms of Eigen-vectors with the highest Eigen-values and these patterns form the feature space for classification. Kernel PCA is a variant of PCA for non-linear feature extraction [Scholkopf 1998].

3.2.2.2 Template Matching

This type of model is widely used in image processing applications to determine the similarity between two pixels, curves or objects. Here, the models of known patterns, known as templates, are available for all classes and the best match is chosen by comparing the test pattern with all available templates [Brunelli 2009]. A measure such as minimum distance or correlation function is then used as a decision variable. The success of pattern classification algorithm depends on the stored templates as well as whether translation, rotation and scale variations of the patterns are taken into account. Such approaches are computationally expensive when the size of template or the image data set increases.

3.2.2.3 Neural Networks

Inspired by the manner in which a biological nervous systems processes the information, Artificial Neural Networks (ANNs) are composed of massively parallel structures of neurons. They are capable of adapting themselves to the data by learning a complex nonlinear input-output relationships. As shown in Figure 3.2, ANNs are organized in different layers, which consists of interconnected nodes containing activation function. The input patterns are presented to the network via input layer, which communicates signals to one or more hidden layers. The signals are processed using a system of weighted connections and the signals are communicated to the output layer, where the classification is obtained.

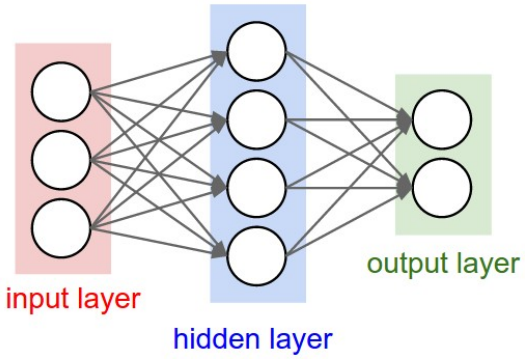


Figure 3.2: Artifical Neural Networks.[1]

The most commonly employed family of ANN for pattern classification is the feed forward networks, which includes multilayer perceptron and radial-basis function networks [Jain 1996].

[1] http://cs231n.github.io/neural-networks-1/

3.2.2.4 Data Clustering

This is an unsupervised approach, in which the aim is to group the data with similar properties into clusters, which can be used for the classification.

Several other classifiers such as support vector machines, decision trees, Bayesian classifiers etc. can be used in the pattern recognition approach. Deep learning based methods are becoming increasingly popular now-a-days, as this technique combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data. Currently, deep learning techniques are state-of-the-art for identifying patterns such as objects in images.

Chapter 4

Sparse Representations and Dictionary Learning

Contents

The advent of digital technology has resulted in generating enormous amount of data. The signals arising from application areas such as remote surveillance, e-commerce, social media, bioinformatics or medical imaging are high-dimensional. For example, in a customer purchase behavior data set, there could be hundreds of thousands of users, each of which is associated with hundreds of products they viewed or purchased. Earth Observation Data is a spatio-temporal data containing thousands of observations which include

geographical and weather information. Several fields have evolved to study the acquisition, processing and classification of such high-dimensional signals. In this chapter, we focus on sparse representations and dictionary learning technique, which has received a special attention over the last few years, for the analysis of high-dimensional signals.

The general principle of sparsity or parsimony is to represent some phenomenon using as few variables as possible. The notion of parsimony is inspired from Ockham's razor, a principle stated by the philosopher William of Ockham, which gives precedence to simple theories over more complex ones. This principle has been incorporated in the fields of statistics and signal processing. In statistics, the models which assume this principle are known as sparse models. They are used in predictive modelling, where the simplest model is selected among several plausible models. In the field of signal processing, the phenomenon is realized through the use of sparse representations, which allows to represent variety of natural signals using linear combination of few basis elements in a set of redundant basis functions. These basis signals can be thought of as a dictionary, with individual basis signals stacked as the columns in the dictionary matrix. Thus, we can represent the signal of length n, with $k << n$ non-zero coefficients. Obtaining sparse signal representation is NP-hard problem, but it can be solved efficiently with greedy algorithms and convex optimization methods. Such high dimensional signals can be reconstructed back by the linear combination of few non-zero sparse coefficients and the corresponding dictionary. In recent years, sparse representation has seen applications ranging from image processing (image deblurring, inpainting, compression etc), speech and object recognition (source separation, classification etc), economics (building models for high dimensional sparse economic data analysis) to bioinformatic data decoding. The algorithms are developed for obtaining sparse representation of data using pursuit methods such as matching pursuit, orthogonal matching pursuit, basis pursuit etc. [Baraniuk 2010, Mallat 1993, Chen 1998]

The choice of the dictionary plays an important role in sparse signal representation. The use of analytic dictionaries such as Wavelets results in the use of predefined basis functions, which have limited data adaptability on account of the fixed mathematical formulation. With the advent of machine learning methods, it became possible to learn the dictionaries from the underlying data. Such approaches are known as dictionary learning methods and they offer greater data adaptability in comparison with the predefined dictionaries. Several methods have been proposed for this task: the method of optimal directions (MOD), K-SVD, sparse K-SVD etc. These methods have been successfully used in image processing applications such as image denoising, restoration, inpainting, compression, classification etc.

In the next sections, we describe the fundamentals of sparse representations and dictionary learning technique, and discuss the most popular methods and applications proposed by the researchers in this domain.

4.1 Sparse Representations

Digital signal can be represented as a weighted sum of Dirac delta functions in time or space. However, this representation does not serve as a good tool for analyzing signals. Many transforms have evolved for the representation of signals using linear combination of fundamental signals known as basis functions. For example, the signal represented as a linear combination of sinusoids gives rise to Fourier representation of the signal and allows the analysis of signal in the frequency domain. Wavelet functions of different translation and scale parameters allow representation of signal in the time-frequency plane. Shifting from the idea of such signal transforms, sparse representations provide a different way of representing the signals.

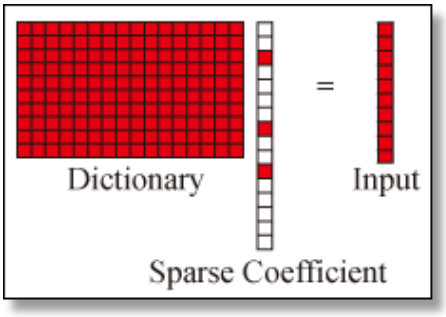


Figure 4.1: Sparse representation of a signal.[1]

Signal modelling using sparse representations consists of describing signal as a linear combination of few basis functions in an over-complete dictionary. The dictionary consists of basis functions or atoms for the representation of the signal. As shown in Figure 4.1, the representation of input signal is possible using few dictionary atoms (marked in red in the sparse coefficient vector).

Consider an over-complete dictionary $D \in R^{N \times K}$. The signal $\mathbf{x} \in R^N$ can be represented as a sparse linear combination of dictionary atoms $\mathbf{x} = D\alpha$. The vector $\alpha \in R^K$ contains the coefficient of the linear combination in representation of the signal $\mathbf{x}$. The representation might be exact $\mathbf{x} = D\alpha$ or approximate, so that $\|\mathbf{x} - D\alpha\|_p \leq \varepsilon$, where ε is the representation error and the norms for measuring deviation can take numerous forms such as l_p norm

[1] http://ranger.uta.edu/~huang/R_Cervigram.htm

with $p = 0, 1, 2$ or ∞. Most of the methods in literature concentrate on the case where $p = 2$. [Aharon 2006]

The sparse representation problem can also be stated as

$$min_\alpha \|\alpha\|_0 , \text{ s.t. } \mathbf{x} = D\alpha \text{ or } \|\mathbf{x} - D\alpha\|_2 \leq \varepsilon \tag{4.1}$$

where $\|.\|_0$ is l_0 norm, counting the number of non-zero entries in the vector. Solving this problem can be stated as finding the sparsest vector α, that represents the original signal $\mathbf{x}$ as a linear combination of columns of dictionary D, and error no more than ε. This process is known as atomic decomposition.

However, minimizing l_0 is a NP hard problem and a common approximation is to replace l_0 norm with l_1 norm. The objective is then to solve the following unconstrained problem

$$\min_\alpha \|\mathbf{x} - D\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{4.2}$$

where λ is called sparsity induced regularizer, which balances the trade-off between reconstruction error and sparsity. This is a convex problem and l_1 constrain induces sparse solutions for the coefficient vector α.

The input signal can be obtained back, with or without loss, by linearly combining dictionary columns with weights indicated by the sparse vector α. This process is referred to as atomic composition. [Elad 2010]

Many algorithms, called pursuit algorithms, have evolved for obtaining the sparse coefficient vector given the signal $\mathbf{x}$ and dictionary D. We briefly discuss several algorithms in the following sub-sections.

4.1.1 Matching Pursuit (MP)

This is a greedy algorithm introduced by Mallat and Zang [Mallat 1993], that optimizes approximations by selecting dictionary atoms sequentially. Given the input signal $\mathbf{x}$ and dictionary D, this algorithm successively finds the dictionary atoms which result in the maximum inner product of signal $\mathbf{x}$ and the indexed dictionary atom. Considering dictionary D formed by n basis functions $D_1, ..., D_n$, the first step is to find D_i that maximizes the inner product of the signal and respective dictionary column D_i.

$$D_i = \arg\max_{D_i \in D} < \mathbf{x}, D_i > \tag{4.3}$$

The corresponding entry in the sparse coefficient vector α_i is set to the inner product. Then, the residual approximation error is given by

$$R_1 = \mathbf{x} - < \mathbf{x}, D_i > D_i \tag{4.4}$$

The algorithm further approximates this residual error by selecting the best dictionary atom in the similar manner described above and iteratively approximates the residual approximations. The process is repeated until a stopping point is reached. The method represents approximate signal using few dictionary columns, when stopped after few steps.

4.1.2 Orthogonal Matching Pursuit (OMP)

This is an extension of matching pursuit method. Here, the dictionary atom is selected only once and all the coefficients extracted so far are updated by computing orthogonal projection of the signal on the set of atoms selected until the corresponding iteration. This results in improvement of the convergence rate. An algorithm can be visualized as selecting the column of D which is most correlated with the present residuals. The sparse coefficient vector is updated and the residual is recomputed by projecting signal $\mathbf{x}$ onto the columns of D that have already been selected. The algorithm iterates until convergence [Pati 1993].

4.1.3 Method of Frames

Considering dictionary vectors as the columns of dictionary D and all sparse approximation vectors as the columns of α, the decomposition of signal $\mathbf{x}$ requires finding solution $\mathbf{x} = D\alpha$. The method of frames selects the one among all solutions of $\mathbf{x} = D\alpha$, for which, the coefficients have minimum l_2 norm. The minimization problem can be stated as $min_\alpha \|\alpha\|_2$, s.t. $\mathbf{x} = D\alpha$. This method is also called minimum length solution as it selects the element of affine subspace containing all the solutions to $\mathbf{x} = D\alpha$, which is closest to the origin. The solution in this case is an average of all possible solutions of $\mathbf{x} = D\alpha$ and is typically of very poor sparsity [Chen 1998].

4.1.4 Basis Pursuit

Chen and Donoho [Chen 1998] proposed a method of decomposition that chooses among many solutions to $\mathbf{x} = D\alpha$, the solution in which the coefficients have a minimum l_1 norm

$$min_\alpha \|\alpha\|_1 , \text{ s.t. } \mathbf{x} = D\alpha \text{ or } \|\mathbf{x} - D\alpha\|_2 \leq \varepsilon \tag{4.5}$$

In exact case ($\varepsilon = 0$), the optimization can be formulated as a linear programming problem, whereas in general case, it takes the form of quadratic problem. There are several efficient solvers for this task and popular among

them are Least-Angle-Regression (LARS) [Efron 2004] and Iterative Shrinkage [Elad 2006a].

Another popular method based on l_1 norm is Lasso. It is different from basis pursuit in the sense that it places restriction on l_1 value, instead of minimizing it. The optimization problem then becomes $min \|\mathbf{x} - D\alpha\|_2$ subject to $\|\alpha\|_1 \leq \lambda$ [Tibshirani 1994].

4.1.5 Focal Underdetermined System Solver (FOCUSS)

The solution for a minimum l_2 norm has a tendancy to spread the energy among large number of entries of α instead of concentrating all energy in few indices. FOCUSS algorithm suggests modification so as to provide a localized energy solution. The solution is found by calculating low-resolution estimate and the sparsity is achieved by pruning process with the use of Affine Scaling Transformation (AST). AST scales the entries of the current solution by those of the solutions of the previous iterations. [Gorodnitsky 1997]

The research in the field of sparse representation is focused on its potential use in many tasks including dimensionality reduction, restoration, compression and classification. In addition to finding the sparse solution using any of the above mentioned techniques, it is the choice of dictionary that forms the crux of signal analysis. The following section briefs about the choice of dictionary, along with the evolution in the dictionary design techniques.

4.2 Dictionaries in Sparse Representation

Depending on the application, signal decomposition techniques change, where the objective is to have a meaningful representation of the signal for capturing the characteristics of the signal. In denoising, for example, the signal representation should isolate noise from the signal of interest. In compression, the signal representation should permit reconstructing signal from small number of feature coefficients, that can be transmitted with less load on the transmission network. Such applications demand decomposition of signal using basis signals so that the signal can be represented as a linear combination of basis elements. Such representation maps the given signal in transform domain, defined by a set of basis functions used for data representation. The operations of thresholding for denoising or discarding coefficients for compression can then be performed in transform domains. A similar approach holds true for classification. Using the fact that significant information in a high-dimensional data lies on a low-dimensional manifold, the features for classification can be

extracted by decomposing signal in different transform domain and carrying out classification on these features, which effectively represent the variation among different classes.

Representing signal thus forms a crucial step in signal processing or classification task. This involves the choice of a dictionary, which is a set of elementary signals called atoms. The signal can be decomposed as a linear combination of dictionary atoms. The choice of fixed dictionaries such as was wavelets, curvelets, contourlets etc. was popular, given the mathematical simplicity and fast numerical computation offered by the approach. They simply "look" at the data as formulated in the basis function design. Wavelet basis with translation and scaling parameters, for example, allows to extract meaningful structures in the data over many scales. Such dictionaries are described algorithmically rather than defining it through an explicit matrix. However, they suffer from the drawback of limited expressiveness. This led to the development of newer over-complete dictionaries, where a dictionary can be formed by combining over-complete set of vectors. In order to obtain the signal decomposition, the basis from an over-complete set of dictionary elements are selected with a sparsity constraint on the representation vector. Such dictionaries allowed to represent wider range of signal phenomena [Rubinstein 2010a, Donoho 2001].

With the arrival of machine learning algorithms, there is a new segment of research which focuses on learning dictionaries from the underlying data itself. The dictionaries are explicit matrix, in this case, and have a property that they are more adaptable to the data. Such finer-tuned dictionaries produce significantly better performance than analytic dictionaries with fixed mathematical formulation. However, such dictionary learning approaches can be computationally demanding, which limits the size of the dictionaries that could be trained and the dimensions of signals that can be processed.

The following subsections throw some light on the research in the dictionary design.

4.2.1 Analytic Dictionaries

As discussed previously, these dictionaries are not explicit and the basis elements can be visualized as parameterized signals. The analytic or fixed dictionaries evolved much later than the popular signal transforms. In the very beginning, the data was seen as linear combination of Dirac delta functions. These functions assumed a value of unity at a single point and zero elsewhere. They hardly provide any usefulness in signal analysis. The famous Fourier basis added great insight in data analysis by transforming the given data into frequency domain, where signal was represented as a linear combination of

sinusoids with different amplitude and frequencies. With the advent of FFT- a faster implementation of Fourier transform, the use of Fourier transform became very popular. However, because of lack of localization, the Fourier transform was not sufficient to analyze non-stationary signals. A slight modification introduced in Short Time Fourier Transform (STFT) saw an immediate application - JPEG image compression. The time-frequency representation of the signal was achieved by applying Fourier transforms over entire duration of the signal. It was assumed that the signal is stationary in the fixed time intervals. STFT was generalized to give rise to Gabor transform, which is a special case of STFT. A Gaussian function is used as a window function before transforming the selected portion of the signal. Further, complex Gabor structures were developed which incorporate directional information and are used in the analysis tasks.

One of the most significant achievement in the field of signal analysis was multi-scale signal representation using Wavelets. The time-frequency representation of signal over multiple scales was achieved using translated and dilated versions of pair of basis signals - scaling function and mother wavelet. Perfect reconstruction filter bank allowed to decompose the signals at multiple levels, unfolding information of signal in different frequency bands and to reconstruct it back using a set of synthesis filters. The theory was formulated in both discrete and continuous domains. The applications of wavelets range in variety of tasks including denoising, compression (JPEG 2000), feature representation etc. Many variants of Wavelet transform such as steerable wavelet transform, stationary wavelet transform, complex wavelet transform were developed further.

At the same time when effective transforms were becoming popular, the sparse representation of signal using few basis elements from a set of overcomplete dictionary was proposed. This triggered the shift from transforms to dictionaries for the sparse representation of signals.

4.2.2 Dictionary Learning

The analytic dictionaries, described in the subsection above, were built by modeling signal by a family of mathematical basis functions. The main advantage offered by this approach is fast implementation. But the dictionary, in this case, can be as successful as its underlying model. With the machine learning techniques rapidly gaining attention, an attempt was made to train the dictionary for obtaining the sparse representation, by extracting information directly from the data. This allowed finer adaption to the complex variations in the data at the expense of increased complexity.

In the dictionary learning approach, the objective is to build a dictionary

D, using given signal $\mathbf{x}$, so that the signal can be represented as a sparse linear combination of dictionary atoms. This is similar to the sparse representation problem discussed as in Equation 4.1, except here, both α and D are to be minimized. The dictionary learning problem can be stated as follows

$$min_{\alpha,D} \|\alpha\|_0 , \text{ s.t. } \mathbf{x} = D\alpha \text{ or } \|\mathbf{x} - D\alpha\|_2 \leq \varepsilon \tag{4.6}$$

Again, several variations of this minimization problem can be reached upon, considering difficulty in solving the non-convex optimization problem above. For example, replacing l_1 norm instead of l_0 norm. Jointly optimizing the sparse coefficients α and dictionary D, however, is a hard problem. Therefore, a two-step iterative process is carried out: (i) In the first step, the sparse coefficients α is fixed and the dictionary D is calculated, and (ii) In the second step, D is kept fixed and α is calculated. These two independent formulations are convex and can be iterated to obtain α and D.

The following subsections describe some of the dictionary learning methods reported in the literature.

4.2.2.1 Method of Optimal Directions (MOD)

This method first uses any pursuit algorithm viz. OMP for finding the sparse coefficient for each signal. The mean square representation error is then calculated as a sum of mean squared differences between each signal component $\mathbf{x}$ and its sparse representation $D\alpha$. In the dictionary update step, the dictionary is updated with an objective of minimizing the representation mean square error obtained in the previous step. The solution for the dictionary update for current iteration is given by

$$D^{n+1} = X\alpha^{(n)T}(\alpha^{(n)}\alpha^{(n)T})^{-1} \tag{4.7}$$

All the dictionary atoms are normalized. The process of calculating the sparse representation and dictionary update is iterated until convergence. The method suffers from relative high complexity in calculation of matrix inversion and several methods have emerged for reducing this complexity. [Engan 1999a, Engan 1999b]

4.2.2.2 K-SVD

This method proposed several modifications in the dictionary update step in the framework used by MOD. The first step to obtain sparse representation of data using any pursuit method remains unaltered. In the dictionary update step, rather than using complex process of matrix inversion, a simple and efficient process is proposed. The dictionary columns are updated one at a

time, keeping others fixed, and the sparse representation vector is updated every time the dictionary is modified. This results in faster convergence than MOD. The name K-SVD is derived from K-means algorithm, which is used as base framework, and Singular Value Decomposition (SVD) approach used for updating individual dictionary elements.

The input to the algorithm is an initial estimate of dictionary matrix, the number of iterations and a set of input signals stacked as columns of input vector matrix $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$. Then, both, sparse representation and dictionary update processes work towards the minimization of common objective function

$$\min_{\alpha, D} \|X - DA\|_F^2 \tag{4.8}$$

subject to $\|\alpha_i\|_0 \leq T$ where $A = [\alpha_1, \alpha_2, ..., \alpha_N]$ and $\|.\|_F$ stands for Frobenius norm, and is given by $\|A\|_F = \sqrt{\sum A_{ij}^2}$.

After obtaining the sparse coefficient vector corresponding to every input signal using N distinct optimization problems, the dictionary atoms are updated sequentially. Considering the sparse representation coefficient matrix $A = [\alpha_1, \alpha_2, ..., \alpha_N]$, the update of dictionary column d_k puts in question only column d_k and the sparse coefficients in k^{th} row. Thus, penalty term can be expressed as

$$\|X - DA\|_F^2 = \left\| X - \sum_{j=1}^{K} d_j A_T^j \right\|_F^2 = \left\| X - \sum_{j \neq k} d_j A_T^j - d_k A_T^k \right\|_F^2 = \left\| E_k - d_k A_T^k \right\|_F^2 \tag{4.9}$$

where A_T^k are the rows of A and E_k is the residual matrix. The problem of finding d_k and A_T^k is tackled by SVD algorithm, but this update process is confined only to those examples whose current representation use the atom d_k [Aharon 2006].

4.2.2.3 Unions of Orthonormal bases

In this interesting approach, the dictionary composed of union of orthonormal bases is considered. Such structure could represent manifolds. This method takes advantage of an efficient pursuit algorithm, known as Block Co-ordinate Relaxation (BCR), for computing the sparse coefficients associated with each orthonormal basis in the dictionary of a union of orthonormal bases.

The dictionary D is a union of L orthonormal bases $D = [D_1, D_2, ...D_L]$, where D_i, $i = 1, 2, .., L$ are orthonormal matrices. The sparse coefficients are represented as $[\alpha_1, \alpha_2, ...\alpha_L]^T$, where α_i contains the sparse coefficients with respect to each orthonormal dictionary D_i. As described in previous

techniques, this dictionary learning approach also uses two steps: coefficient update and dictionary update.

In the first step, the coefficient matrix is found using BCR. The overall minimization problem is split into L steps, one for each of L layers, while keeping all other components of sparse matrices fixed. In the dictionary update step, the orthonormal basis D_i are updated one-by-one. First, the residual error is calculated as $E_i = X - \sum_{j \neq i} D_j \alpha_j$. The SVD of the matrix $E_i \alpha_i^T = U \Lambda V^T$ is then calculated and the ith orthonormal basis is updated as $D_i = UV^T$ [Lesage 2005, Rubinstein 2010a]

4.2.2.4 Sparse Dictionaries

This method to generate a dictionary that combines the advantages of analytic dictionaries and those learned from the data was presented recently. They observed that the dictionaries learned using K-SVD is highly structured, with noticeably regular atoms. The sparse dictionary model was therefore proposed, which suggested that each atom of the dictionary has itself a sparse representation over some prespecified base dictionary B. Thus, the sparse dictionary, itself, can be decomposed as

$$D = BA \tag{4.10}$$

where A is the atom representation matrix and its sparse nature gives rise to sparsity of each columns of D. Such dictionary model, when compared to analytic dictionaries, provides adaptability via modification of the matrix A as well as choice of implicit dictionary B. When compared with explicit dictionary, it provides more efficient and compact sparse structure for storage and transmission, and requires less number of samples for training the dictionary.

For learning a dictionary, an algorithm known as sparse K-SVD was proposed. It is an extension of already developed algorithm K-SVD and involves two steps of coefficient updates and dictionary update. The application of algorithm for generalization and denoising of CT volume data has been presented and it has been shown that the algorithm improves generalization. The training method can be used to learn larger dictionaries, for example, large image patches or 3D image patches [Rubinstein 2010b].

4.2.2.5 Online Dictionary Learning

In case of large-scale dataset with huge number of training samples, the objective function minimization in the dictionary learning formulation poses computational challenge. Previously proposed methods access the whole training data at each iteration to solve some minimization problem and could not handle very large training data. Online dictionary learning approach, based on

stochastic approximations, processes one element or a smaller subset of the data set for faster convergence.

Considering training samples as i.i.d. with distribution $p(\mathbf{x})$, the elements are collected one at a time and the steps of sparse representation (using LARS-Lasso algorithm) and dictionary updates (using block coordinate descent) are carried out. The minimization function used for dictionary update agglomerates the computations from previous iterations and acts as a surrogate function for the empirical cost function. The convergence speed can be further improved by drawing multiple signal examples at each iteration, an approach known as mini-batch extension. [Mairal 2009b]

4.3 Dictionary Learning in Classification

In this section, we discuss dictionary learning methods in classification. The conventional dictionary learning methods, described above, are not optimal for classification as they are simply used for signal representation. For enabling the use of dictionary learning in classification, several approaches have been proposed to learn the classification oriented dictionary, in a supervised setup. These methods either use representation error or sparse coefficients for performing classification, where dictionaries are forced to be dicriminative. The following subsections describe few such approaches.

4.3.1 Sparse Representation Based Classification

This method [Wright 2009] reports the use of discriminative nature of sparse coefficients in face recognition. The basis elements in the dictionary are the original face images in the training set. Given the sufficient number of training samples from each class, the test image will be represented as a linear combination of few samples from the training samples. Therefore, the classification is performed by seeking the sparsest representation, which automatically discriminates between various classes.

Given C classes of individual faces, the dictionary is represented as $D = [X_1, ..., X_C] \in R^{d\mathbf{x}N}$, where $X_c \in R^{d\mathbf{x}N_c}$ is a subset of N_c individual faces which belong to class c. Then, for a query image $\mathbf{y}$, the method finds the sparse representation α over dictionary D via l_1-norm minimization

$$\alpha = \arg\min_{\alpha} \ \|\alpha\|_1 \ \text{ subject to } \ \|\mathbf{y} - D\alpha\| < \varepsilon \tag{4.11}$$

In the subsequent step, the test image $\mathbf{y}$ is assigned to class c such that

$$c = \arg\min_{i} \ \|\mathbf{y} - D\delta_i(\alpha)\|_2 \tag{4.12}$$

where $\delta_i(\alpha)$ has non-zero entries at indexes associated with class c.

This method achieved impressive results for face recognition and was robust to occlusion and lighting.

4.3.2 Meta-Face Learning

With the previous approach, using entire training data as a dictionary becomes computationally expensive for calculating sparse coefficients, if the size of the training data is large. Furthermore, the training images in face recognition example have redundancy and noise, which could degrade the classification. This method [Yang 2010b] learns a more compact and robust set of bases, which are called metafaces, by learning a dictionary for individual class. Given the data samples X_c for class c, the dictionary D_c is obtained for that class. The class specific dictionary learning formulation can be represented as

$$\min_{\alpha_c, D_c} \|X_c - D_c\alpha_c\|_2^2 + \lambda \|\alpha_\mathbf{c}\|_1 \tag{4.13}$$

Then, individual class dictionaries are concatenated to form an overall dictionary $D = [D_1, ...D_C]$ and the classification is performed in the similar manner as described in the sparse representation based classification method above.

4.3.3 Dictionary Learning with Structured Incoherence

Ramirez et al observed that the sub-dictionaries for each class may share common bases and can be coherent. Interchangeable use of such coherent dictionary atoms in the calculation of sparse coefficients for the test image could introduce errors in the reconstruction error-based classifier. This method [Ramirez 2010], therefore, introduced an incoherence promoting term to make the dictionaries associated with different classes as independent as possible.

$$\min_{\{\alpha_c, D_c\}_{i=1,2.,..C}} \sum_{i=1}^{C} \left\{ \|X_c - D_c\alpha_c\|_2^2 + \lambda \|\alpha_\mathbf{c}\|_1 \right\} + \eta \sum_{i \neq j} \left\| D_i^T D_j \right\|_F^2 \tag{4.14}$$

The addition of incoherence term $\left\| D_i^T D_j \right\|_F^2$ minimizes the coherence between dictionary atoms of different classes. It was observed that, even after adding incoherence, atoms representing common structures in all classes appear in the sub-dictionaries and are often used in the sparse reconstruction coefficients. This would make the reconstruction costs similar and degrade the classification. An improvement was suggested by ignoring the coefficients associated with these common atoms as the reconstruction coefficients having high absolute value.

4.3.4 Fisher Discrimination Dictionary Learning (FDDL)

Previous methods only uses the reconstruction error for each class as an information for classification, whereas the sparse coefficients are not discriminative. Yang et al. [Yang 2011] introduced the Fisher criterion into the dictionary learning framework, which forces the inter-class sparse coeffiecients to be discriminative. This subsequently propagates the discriminative power to the class specific dictionaries as well as the sparse coefficients. The reconstruction error and the sparse coefficients can therefore be used in the classification step.

This method introduces two additional terms in the conventional dictionary learning objective function: A discriminative fidelity term and a discriminative coefficient term.

$$\min_{\alpha,\mathbf{D}} R(X, D, \alpha) + \lambda_1 \|\alpha\|_1 + \lambda_2 f(\alpha) \tag{4.15}$$

where $R(X, D, \alpha)$ is the discriminative fidelity term, which ensures that each sub-dictionary corresponding to each class has good representation power to the samples from the same class, but has poor representation power to the samples from all other classes.

The second term $\lambda_1 \|\alpha\|_1$ introduces sparsity constraint, where as the last term $\lambda_2 f(\alpha)$ is the discriminative coefficient term that makes the coding coefficient of X over D discriminative. This is achieved by using Fisher Criterion, which minimizes the within-class scatter and maximizes the between-class scatter of sparse coefficients α.

4.3.5 Discriminative K-SVD

As proposed by Zhang and Li [Zhang 2010a], discriminative K-SVD is an extension of K-SVD method, which adds discrimination power into the dictionary by introducing a linear classifier in the conventional dictionary learning objective function. The introduction of the classification error from the linear classifier in the objective function results in finding the best data representation dictionary and solving for the classifier, simultaneously. An application on face recognition is demonstrated to validate the method.

4.4 Applications of Dictionary Learning

Dictionary learning allows learning basis functions for the representation of signals on the fly, as apposed to using pre-defined basis which are assumed to

be general enough to represent the signal. Over the past few years, many researchers have demonstrated that sparse representations and dictionary learning can achieve state-of-the-art results in many applications.

These approaches can be classified into two categories: dictionary learning for data representation and classification. In the first set of applications, the objective is to learn dictionaries which better represent the data. The applications, in this category, include denoising [Li 2012, Elad 2006b], image super-resolution [Yang 2010a], image inpainting [Mairal 2008a] and image compression [Bryt 2008].

In the second category, the ability of dictionary in data discrimination is as important as data representation. The classification approaches are developed by learning dictionaries which promote the data discrimination. Some of the applications include audio classification [Grosse 2012], texture classification [Mairal 2008b] and face recognition [Wright 2009, Zhang 2010a].

4.5 Summary

Sparse and redundant representations provide means to decompose data using set of basis functions or dictionary. Many theories and algorithms have evolved for atomic decomposition and dictionary learning. The applications either use analytic dictionaries or those learned from the data. Current research focuses on designing efficient methods for dictionary learning for high dimensional and large data sets and classification of such data.

Chapter 5

Role of Dictionary Size in Pattern Classification

Contents

Sparse representation allows signals to be represented with as few variables as possible. The selection of basis vectors, which are in turn used for obtaining sparse coefficients for representing the signal is an important task. As published by Olshausen et al. [Olshausen 1996, Olshausen 1997], a set of basis functions or a dictionary can be learned from underlying data. The dictionary learning, instead of using fixed off-the-shelf dictionaries, offers better data adaptability and has led to many successful applications in the field of image processing [Raina 2007, Song 2012, Elhamifar 2012]. These approaches rely on the fact that natural signals and images have predominant lower-dimensional structure and can be represented using a few or sparse coefficients. In applications such as image denoising [Elad 2006b], image restoration [Mairal 2009a] and image super-resolution [Yang 2010a], the dictionaries are learned mainly for data-representation. The success of these methods lies in the representational power of the learned dictionaries. The dictionaries

learned in this manner results in lower reconstruction error and the dictionary columns represent the best basis functions adapted to the given data set. However, in other applications such as image classification, only data representation might not be enough.

The dictionaries, when used in classification, need to have discrimination power in addition to having a good representation capability. Many approaches have been proposed for discriminative dictionary learning. These approaches learn the dictionaries that are suited for data representation as well as discrimination between class data. These methods can be sub-categorized in several ways.

- Most of the approaches modify the objective function used in the classical dictionary framework so that a part of the objective function assures data representation ability while the other part encourages discrimination between class data. Such approaches fall under category of supervised learning. However, there are also unsupervised methods that use dictionary learning for classification.

- In supervised learning, there are some methods which learn a single dictionary that holds the discrimination information, where as several other methods that learn separate dictionaries for individual classes, are then used for classification.

- Some discriminative dictionary learning methods use image themselves as the basis functions in the dictionary. Such methods are primarily used where global classification of image is under consideration. With higher dimensionality of the input images, the computational complexity can pose issues in such methods. On the other hand, there exist other methods, which use image blocks that subdivide image into overlapping patches or volumes. The dictionaries are learned using these sub-blocks and the classification is obtained on sub-block level, before global classification is obtained.

- The discriminative dictionary learning methods can be used for the classification of overall images such as face recognition, image categorization or they can be used for local image analysis and segmentation like applications.

- The discriminative dictionary learning methods typically use reconstruction error for the classification. However, there are some methods which make use of sparse coefficients along with the reconstruction error.

A popular dictionary learning method for better image representation (as seen in Chapter 4) is KSVD [Aharon 2006], which learns an over-complete

dictionary from the training data set of image patches. It is not suited for optimal classification as the objective of this method is to learn the dictionaries that are better suited for data-representation tasks. Yang et al. [Yang 2010b] proposed a method for face recognition, where the dictionaries are learned for individual classes and the concatenated dictionary is then used to obtain sparse coefficients and reconstruction error for the given test image, based on which, the classification is achieved. Ramirez et al. [Ramirez 2010] introduced an incoherence promoting term in the classical dictionary learning formulation so that the dictionaries learned for different classes are as independent as possible. Even after adding incoherence, they note that the dictionary atoms representing common features for all classes are used frequently in deriving sparse coefficients and this gives rise to higher absolute value for the corresponding coefficients. Therefore, they proposed to discard such sparse coefficients in calculating reconstruction error, which is finally used as decision variable for classification. Mairal et al. [Mairal 2008b] proposed the supervised dictionary learning method by introducing the logistic loss function in the conventional dictionary learning framework and validated their method using digit recognition and texture classification. Zhang et al. [Zhang 2010a] extended K-SVD method by incorporating the label information in the dictionary-learning stage, which adds discrimination information. The method is verified using commonly used face recognition data sets such as YaleB [Georghiades 2001] and AR database [Martínez 1998]. Label consistent K-SVD proposed by Jiang et al. [Jiang 2011] contains label information as mentioned in previous approach and in addition, a discriminative sparse-code error term is introduced to force the signals from the same class to have a similar sparse coefficients. Yang et al. [Yang 2011] proposed Fisher discrimination dictionary learning method. They imposed Fisher discrimination criterion on sparse coefficients to make them discriminative and discussed its applications in digit recognition, gender classification and face recognition.

The discriminative dictionary learning algorithms discussed above modify the dictionary learning objective function in such a way that the discrimination information is added into the learned dictionaries, which are then used for the classification. However, there are few other approaches which use the standard dictionary learning formulation in the classification. Ren et al. [Ren 2015] and Weiss et al. [Weiss 2013] use dictionary learning in unsupervised manner to detect abnormal events or multiple sclerosis lesions as outliers. In these approaches, the dictionaries are learned to capture the global trends in the given data set. The atoms of the dictionaries represent a particular normal behavior, whereas the rarely occurring events or outliers are differentiated from the data exhibiting normal behavior using reconstruction error obtained from these dictionaries. There are few other

approaches that isolate the procedure of classification from the dictionary learning [Zhang 2009, Mairal 2008c, Rodriguez 2007]. In these approaches, the dictionaries are learned first and the features extracted using these dictionaries are fed to a classifier like SVM in order to achieve classification.

There are several drawbacks associated with the discriminative dictionary learning techniques discussed above: (a) The main drawback of these methods is the computational complexity introduced by additional terms in the dictionary learning problem. Owing to the large time requirements, the discriminative dictionary learning methods limit their usage only when high computing solutions are available. (b) In some discriminative dictionary learning problems, the objective function is non-convex and the solutions for updating dictionary columns and the sparse coding does not guarantee global minimum. (c) Often, additional parameters are introduced in such methods. It is very difficult to tune these parameters for a particular application. Experimenting with more number of parameters essentially shifts the focus of research to parameter tuning and as one set of experiment requires a large amount of time, finding parameters using grid method might take enormous time. (d) Many of these methods have been proposed for tasks such as face recognition or texture classification. However, when these methods are to be incorporated in different application and the experimental results are not as expected, it is difficult to trace back and conclude why the method does not work on specific applications.

There are two very important parameters in the conventional or standard dictionary learning framework, namely the sparsity parameter λ and the dictionary size L. It is well known that the sparsity parameter controls the portion of non-zero coefficients participating in sparse decomposition vector as compared to the number of available atoms in the dictionary. Higher the value of λ, the lesser number of non-zero values are favored in the sparse representation vector. The effect of this parameter in penalizing sparse solution is well studied and experimented [Tibshirani 1994]. However, the role of dictionary size in image classification has not been much explored yet in the signal processing and machine learning community. In this thesis, we carried out a detailed study of how dictionary size affects the classification and showed that this parameter is crucial in image classification, as described next.

5.1 Why is Dictionary Size Important?

As described earlier, the conventional dictionary learning framework is focused on how well data can be represented with a sparsity constraint on the data representation vectors. The choice of basis functions, which are arranged as

the columns of the learned dictionary, plays a vital role in the data representivity. The number of such basis functions, and hence the dictionary size thus directly controls the set of basic building blocks that are used to obtain the sparse coefficients and consequently represent the given data. Using very few dictionary atoms might result in under-representation of the data, whereas a large number of dictionary atoms might capture the detailed structures within data set and thus result in over-representation of the data. While over-representation of data is acceptable in applications such as denoising or compression, it can lead to detrimental performance in classification, thus making the dictionary size a very important parameter.

For obtaining classification using dictionary learning approaches, generally the dictionaries are learned for each class and the reconstruction error obtained using these dictionaries is compared to find the best representative dictionary. However, the comparison of reconstruction errors would be meaningless if there is relative under- or over-representation of class data using the dictionary for the corresponding class. The dictionary size or the number of basis functions control the data representation power of the dictionaries. One way to improve the classification is to select the dictionary size that leads to having the same level of representativity for all classes. Thus, choosing the correct dictionary size can avoid the relative over or under-representation of class data and hence improve the classification accuracy.

The role of dictionary size becomes even more significant when there are differences in variability of the class data. The patterns of interest in a given data set belong to one particular class, as opposed to the background data representing the opposite class. These patterns might include less occurring or relatively smaller structures in the image, whereas the background data, on the other hand, might involve some more complex structures and have more variability when compared with the data from other class. To illustrate this, let us consider an example of activity recognition in airport surveillance video. To detect activities like person talking on mobile or picking a bag, it can be observed that the background is associated with higher variability as compared to the activities of interest as it contains more complex information. If this data is to be represented in terms of different classes, we can learn several dictionaries for each class. However, the class specific dictionaries of same size would not consider the variability differences between class data. Different classes in this application do not have the same variability. Therefore, learning dictionaries of the same size could result in best possible data representation for each class, but the dictionaries learned in this manner might not be effective in performing the classification.

The effects of dictionary size in applications such as image categorization are only briefly discussed previously in [Gao 2014]. The size of the dictionary

might not be as significant in applications considered so far, such as face recognition. The reason behind this is that there is not much variability difference between face images of two persons. However, if we consider an application to detect lips or eyes from face images, there is a shift in the level of variability between class data. The dictionary size to represent both classes of data could drive the classification results. This phenomenon might become important in the classification of body structures or pathology instances in medical image data sets such as detection of brain pathologies. The background structure could possess huge variability when compared with the patterns of interest and the role of dictionary size in terms of data representation as well as differentiation could be significant.

In this work, we analyzed the role of dictionary size in image classification. The main idea behind our work is to explore the variability differences between class data and study how the size of the dictionaries for each class could be tuned, in order to achieve better classification. We also studied the discriminative dictionary learning algorithm such as Fisher Discrimination Dictionary Learning (FDDL) [Yang 2011], and studied the significance of dictionary size in this discriminative dictionary learning technique. The following sections describe the dictionary learning based pattern recognition in applications such as handwritten digit recognition and lips detection in face images, whereas the next chapter describes the work on brain pathology detection in multi-channel MR images and the significance of dictionary size in more complex medical imaging application.

5.1.1 Significance of Dictionary Size with Example on USPS Handwritten Digit Database

Consider an example of the United States Postal Service (USPS) database [1], which consists of the handwritten images of digits from 0 through 9. We develop the dictionary learning based classification method to correctly identify unseen images in this data set and study the significance of the size of the dictionaries used for the classification.

The USPS database contains 9298 grayscale images of size 16×16. The training data set consists of 7291 images of digits 0 through 9. The number of images for each digit are not the same and their numbers vary from 500 to 1200. The test data set consists of total 2007 images for digits 0 through 9, with the number of images for individual digits varying from 150 to 350. These number indicate that there is class imbalance within the data set. Few examples of digits are shown in Figure 5.1. It can be seen that there are

[1] http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html

variations in the manner in which the same digit is written and this makes the classification task more difficult.

Figure 5.1: USPS data set: Training examples

We extended our training data set by translating each image up, down, left, right, up-left, up-right, down-left, down-right. To classify the test images, we designed the classifier as mentioned in the following sub-sections.

5.1.1.1 Image Normalization

The images for each digit are normalized so that each image corresponding to individual digit has a unit l_2 norm. The images are then flattened to form one-dimensional vectors for each image.

Let X_i denote the training data matrix for digit i, where each column is one training sample of the corresponding digit. The overall training matrix is indicated as $X = [X_1, X_2, ..., X_c] \in R_n^d$ where c is the number of classes $c = 0, 1, .., 9$, $d = 256$ is the dimensionality of each input image and n is the total number of training samples for all classes. Let $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m] \in R_m^d$ indicate the test data matrix with $m = 2007$ digits in the test data set.

5.1.1.2 Dictionary Learning

In this step, the dictionaries are learned for each class $c = 0, 1, .., 9$, using the training data X_c for the corresponding class. We used the online dictionary learning algorithm, as described in Section 4.2.2.5, which is significantly faster than batch alternatives on large dataset [Mairal 2009b]. Let $D_1, D_2, ..., D_c$ be the dictionaries for the classes $c = 0, 1, .., 9$. The data for each class can be represented by the dictionary for the corresponding class, whereas the dictionary for other class will not faithfully reconstruct the data from all other classes. Thus, we can use the dictionaries obtained in this manner for classifying the test image.

5.1.1.3 Sparse Coding

Given a test image $\mathbf{y}$, the classification is a two-step process. In the first step, we calculate the sparse coefficients α_c for each class $c = 0, 1, .., 9$, by solving the following optimization problem

$$\min_{\alpha_c} \|\mathbf{y} - D_c\alpha_c\|_2^2 + \lambda \|\alpha_c\|_1 \tag{5.1}$$

5.1.1.4 Classification

The sparse codes α_c are the representation coefficients for the signal $\mathbf{y}$, using the class dictionaries D_c. We can thus assign the test image to class with the minimum reconstruction error as given below

$$\operatorname*{argmin}_c \|\mathbf{y} - D_c\alpha_c\|_2^2 . \tag{5.2}$$

Using sparsity parameter $\lambda = 0.95$ for which the best results were obtained and the dictionary size of 255 for all classes, we achieved an error rate of 3.44%. Traditional dictionary learning based classification methods use the same dictionary size in their approaches. However, this does not take one very important aspect into consideration: The variability differences between class data. The dictionaries are learned in order to achieve the best reconstruction, however, this might not be sufficient in the case of classification. The use of the same dictionary sizes for each class could lead to relative over- or under-representation of the class data. The relative representation power of the dictionaries is important in the classification tasks, where comparison of reconstruction error is performed in order to decide the class label. The dictionary size is the parameter that can control the relative representation for each class and subsequently result in better classification.

It can be clearly seen from the images of all digits that the digit 1 has less complexity than the rest of the digits. Therefore, we can allow the same dictionary size of 255 for the digits except 1 and lower dictionary size, for example 100, only for the digit 1. The error rate, using these dictionary sizes, reduces to 3.34%.

The improvement in the classification using dictionary of different size as compared to the classification obtained using dictionaries of same size indeed suggests that the dictionary size plays a major role in image classification. More experiments with different set of dictionary size could further improve the result, but we provided this simple example only to demonstrate that the dictionary size can be adapted according to the complexity of the class data in order to achieve better classification. We build upon this concept to investigate in more detail how can we choose the dictionary size for each class.

5.2 Dictionary Size Selection

5.2.1 Methods

In the last section, we described the motivation behind the use of different dictionary sizes while developing a pattern recognition application using sparse representation and dictionary learning framework. We demonstrated the significance of dictionary size in a simple application such as handwritten digit recognition. In this section, we discuss various methods for estimating the dictionary size for each class in pattern recognition applications where there are variability differences between class data. We further illustrate these methods for pattern classification application such as lips detection in face images.

5.2.1.1 Dictionary Size Selection using PCA

In the dictionary learning formulation, the objective function is defined for achieving good data representation, in addition to the sparsity constraint on the data representation coefficients. In practice, the selection of large enough dictionary size achieves good data representation, but using the dictionaries of the same size for individual classes for classification might not be a good idea. While such dictionaries are good for data representation, they might not be suited for better classification. In this method, we propose to use the principal component analysis (PCA) of the training data to select the dictionary size for each class based on the complexity differences between class data. The main motivation behind this idea is that the patterns of interest in many applications are often less complex structures or a less occurring phenomenon in a relatively complex background. PCA can be used to capture these variability

differences and the dictionary size for each class can be adapted based on the PCA of the data for the corresponding class.

PCA is widely used for dimensionality reduction, feature extraction, lossy data compression and data visualization. The main idea behind PCA is to orthogonally project data onto a lower-dimensional linear space, also known as principal subspace, so that the variance of the projected data is maximized. It transforms the data into lower-dimensional subspace where few principal components explain the maximum amount of variance in the data. The principal components are selected in incremental fashion by choosing each new direction to be that which maximizes the projected variance amongst all possible directions orthogonal to those already considered. If we consider M dimensional projection space, the optimal linear projection for maximum variance of the projected data is defined by the M eigenvectors of the data covariance matrix corresponding to the M largest eigenvalues.

We selected a specific value of cumulative variance, for example 95% or 98%, and recorded the number of eigenvectors required to attain the target value of cumulative variance for each class. The number of eigenvectors, obtained in this manner, for each are indicative of the complexity information for the corresponding class data. The selection of dictionary size for each class based on the relative number of principal components or eigenvectors for the corresponding class should consider the variability differences between class data. The use of same dictionary size for all classes ignores the complexity differences between class data and hence, the classification achieved with the use of same dictionary size for all classes might not be optimal.

However, the success of this method depends on the probability distribution of data under consideration. PCA applies well if the data is linear or if the underlying distribution is Gaussian. PCA fails in the cases where the data is non-linear and in such cases, adapting dictionary size based on the variability analysis using PCA, as described above, could lead to errors.

5.2.1.2 Dictionary Size Selection using Histogram based Measures

In the previous subsection, we discussed the dictionary size selection using PCA. Given the training data for each class, the PCA of data provided a direct measure of complexity for each class. The dictionaries learned from the training data were not involved in this analysis. A different, but also natural, approach would be to compare the representation power of dictionaries learned for each class. The mean reconstruction errors for each class data using the dictionary learned for the corresponding class could be one such measure, which would describe how well the dictionaries are representative for each class data. For class specific dictionaries of the optimal different size,

the similar values of the mean reconstruction errors for each class data would be suggestive of the same level of representativity using each class dictionary. We computed this measure for different dictionary sizes for each class data, but failed to find any conclusive evidence on how dictionary size could be selected using this measure. This led us to investigate the histogram of reconstruction errors as a tool for comparing the representation power of the learned dictionaries, as described next.

Consider a two-class classification problem. The dictionaries learned for individual classes achieve lower reconstruction error for the corresponding class data. Let $\mathbf{x}_c^i$ indicate n data samples for the class $c = 1, 2$ and $i = 1, 2, .., n$. Denote the dictionaries for individual classes as D_c. We can learn the dictionaries from the training data and calculate the sparse coefficients α_c^i by solving the optimization problem below

$$\min_{\alpha_c^i} \left\|\mathbf{x}_c^i - D_c\alpha_c\right\|_2^2 + \lambda \left\|\alpha_c\right\|_1 \tag{5.3}$$

The reconstruction error for individual data sample i, using the dictionary D_c can be given by

$$R_c^i = \left\|\mathbf{x}_c^i - D_c\alpha_c^i\right\|_2^2 \tag{5.4}$$

We then calculate the histogram of reconstruction errors for a class data, using the dictionary learned for the corresponding class. Each signal in the given class is said to be faithfully represented by the dictionary learned for that class. This is important in the classification strategy, as the learning model must capture the trends in the training data and data representation ability of the dictionary indicates how well the model has learned from the underlying training data. Now, if the size of the dictionary for a particular class is too small, the dictionary might have limited data representation capabilities. On the other hand, if a very large dictionary size is used in the dictionary learning formulation, the dictionary might capture every minute detail within the training data and this might lead to over-representation of one of the class data. One of the ways to guarantee that our model or dictionaries do not cause the relative under- or over-representation is to have the same level of representativity for both the classes with the use of individual dictionaries. The histograms of reconstruction errors provide one way to measure how well the dictionaries represent each class data and matching these histograms could gurentee the same level of representativity using dictionaries for each class.

Using sufficiently large dictionary size for one of the classes, we keep the dictionary size for this class constant and obtain the histogram of reconstruction errors for the given class data. Our objective is to select the dictionary size for the second class, from several possible dictionary sizes. Lower dictionary size for this class will result in under-representation of the class data relative

to the first class, which will lead to the higher reconstruction errors. This will subsequently result in the mis-match of two histograms that correspond to the reconstruction errors of each class data. Similarly, large dictionary size for the second class will lead to over-representation of this class data as compared to the data for the opposite class and this will also result in histogram mis-match. The optimal dictionary size can thus be selected by varying the dictionary size for the second class and comparing the histogram of reconstruction errors for this class data with the similar histogram obtained for the first class. The dictionary size for which the two histograms match each-other is selected as the optimal choice.

Let $H^m_{1,1}$ and $H^m_{2,2}$ denote the histograms obtained using the reconstruction errors $R_{1,1}$ and $R_{2,2}$ respectively for the class data X_1 and X_2, and the dictionaries D_1 and D_2 respectively, where m is the number of bins in the calculation of histograms. For the comparison of two histograms, we use the Jeffreys divergence metric calculated as follows

$$d_{J1,L}(H_{1,1}, H_{2,2}) = \sum_m (H^m_{1,1}\ log\frac{H^m_{1,1}}{t^m} + H^m_{2,2}\ log\frac{H^m_{2,2}}{t^m}) \tag{5.5}$$

where $t^m = \frac{(H^m_{1,1}+H^m_{2,2})}{2}$ and L is the dictionary size for the lips class.

Jeffreys divergence d_J is an improvement over Kullback-Leibler (K-L) divergence d_{KL}, a popular measure of similarity between two probability distributions. While K-L divergence is not symmetric, the Jeffreys divergence provides a symmetric measure of similarity. The smaller value of this metric indicates more similarity between two histograms a and b .

$$d_J(a, b) = d_{KL}(a, b) + d_{KL}(b, a) \tag{5.6}$$

However, for classification, we also need that the dictionary for one class should not be representative of the data from the opposite class. Thus, we can use the similar idea as discussed above and compare the histograms of the reconstruction errors obtained for the class data using the opposite class dictionary. The matching of these two histograms would suggest that the data for each class is equally badly represented by the dictionary for the opposite class. Therefore, among several dictionaries with different sizes, we select the dictionary size which results in matching of histograms for the reconstruction errors of the class data derived from the other class dictionaries. We fix the dictionary size for one class and select the dictionary size for the other class.

Let $R_{1,2}$ and $R_{2,1}$ denote the reconstruction errors for the class data X_1 and X_2 obtained using the dictionaries of opposite classes D_2 and D_1, respectively. Let $H^m_{1,2}$ and $H^m_{2,1}$ denote the histograms obtained using these reconstruction

errors, where m is the number of bins in the calculation of histograms. The Jeffreys divergence for comparing these two histograms is given by

$$d_{J2,L}(H_{1,2}, H_{2,1}) = \sum_m (H^m_{1,2}\ log\frac{H^m_{1,2}}{u^m} + H^m_{2,1}\ log\frac{H^m_{2,1}}{u^m}) \tag{5.7}$$

where $u^m = \frac{(H^m_{1,2}+H^m_{2,1})}{2}$.

For several dictionary sizes, the minimum values of the Jeffreys divergence measures $d_{J1,L}(H_{1,1}, H_{2,2})$ and $d_{J2,L}(H_{1,2}, H_{2,1})$ would thus suggest the same relative behavior in representing both class data using dictionaries for the same and the opposite class, respectively. We, therefore, consider the squared sum of these measures and select the optimal dictionary size for which the value of this measure is minimum.

$$\underset{L}{\text{argmin}}\ d^2_{J1,L}(H_{1,1}, H_{2,2}) + d^2_{J2,L}(H_{1,2}, H_{2,1}) \tag{5.8}$$

5.2.1.3 Dictionary Size Selection using Empirical Method

As described in the previous subsections, the dictionary size of each class is a crucial parameter in pattern recognition applications and the analysis of training data or the dictionaries learned from the training data could suggest the optimal dictionary size for each class better suited for classification. In this section, we describe the selection of dictionary size using empirical method, where the values of the size of the dictionaries for each class can be found experimentally. The difference in variability of the class data is explored to decide the dictionary size for each class. The optimal dictionary size among dictionaries of various sizes is selected by performing classification on the training data and the optimal values of dictionary size chosen from this experiment are then incorporated for the validation on test data.

In this method, the given data set is divided into training and test set. The dictionaries are then learned using the training data and the classification is first performed on the training data itself. A fixed dictionary size is selected for the class associated with higher variability and the dictionaries of various size are then learned for the opposite class. The traditional methods use the same dictionary size for both the classes, however, our hypothesis is to use different dictionary sizes for each class to take into account the differences in variability of the class data. A classification measure such as Dice-score is chosen to compare these classification models, each corresponding to the different dictionary size for the class having lower variability, and pick the best model among them. The classification performed on training data thus gives the optimal dictionary sizes for each class. In the next step, the dictionaries selected in the previous step are employed for the classification of the test

data and the results of classification are validated. As we will show later, the optimal ratio between dictionary size for different classes is quite constant between the training and the test data sets. Therefore, if we find the optimal ratio of dictionaries with a fixed size for one class, it is still useful afterwards.

5.2.2 Experiments and Results

We selected the application of lips detection in face images as it provides a typical computer vision example, where the dictionary learning methods are usually evaluated, and is also a good illustration of the problem under consideration. The lips are associated with less variability when compared with more complex face structures other than lips, combined together.

We used PUT Face database [Kasiński 2008], which consists of 9971 face images of 100 persons in partially controlled illumination conditions over uniform background with different pose variations. Several images for each person were captured in the following series: The head turning from left to right, the head nodding from the raised to the lowered position and few images without any constrains on the pose. The example images for three persons, each with three different poses are shown in Figure 5.2. Each row indicates the poses of the same person in the database.

All images in the database are manually annotated for a face, eyes, nose and mouth or lips. The rectangles defining each of these facial structures are provided along with the image database.

We divided the data set by randomly selecting 70 persons for training and 30 persons for testing. To reduce the computational complexity for the dictionary learning algorithms, we randomly selected 3 poses for each person and resized the images to 512×512. We then restrict further analysis to the face region, with the use of the face annotations. The image patches of size 15×15 in the face region are extracted and labelled as either lips or non-lips class. For the purpose of labelling the patches, we used a predefined threshold of 80%. If the number of pixels that belong to the mouth annotation in the image patch under consideration are greater than this threshold, the patch is labelled as lips patch. Otherwise, it is labelled as non-lips patch. Such labelling resulted in around 10K patches for the non-lips class and few hundreds of patches for the lips class, for each pose of a person.

We learned the dictionaries for the lips and the non-lips class, using the training data, as described in Section 5.1.1.2. Given the test data, the dictionaries learned are then used to obtain the sparse coefficients and the test patch is assigned to the class corresponding to the dictionary with minimum reconstruction error. However, the selection of dictionary size still remains an important issue, as described in the previous subsection. We describe the

	95%	98%	99%
Non-lips	30	60	92
Lips	18	39	65

Table 5.1: Principal component analysis of the training data for the lips and the non-lips class. For each class mentioned in a row, an entry in the table denotes the number of eigenvectors required to attain the percentage of total variance indicated in each column.

experiments and results for the dictionary size selection for the lips detection application in the following subsections.

5.2.2.1 Dictionary Size Selection using PCA

We performed PCA on the training data for the lips and non-lips class, after normalization. Each vector of the training data was normalized for unit l_2-norm. The PCA of the data then gives the most predominant vectors, also known as principal components, associated with the largest possible variance of the underlying data.

Table 5.1 shows the number of principal components required to reach 95%, 98% and 99% of cumulative variance for each class: the lips and non-lips. It can be seen that the number of eigenvectors required for the representation of non-lips class are approximately 1.5 to 2 times the number of eigenvectors for the lips class. Figure 5.3 shows the variation of cumulative variance with the number of eigenvectors for the lips and the non-lips class. The data corresponding to the lips class achieves the percentage cumulative variances mentioned earlier using less number of eigenvectors than the non-lips class. This suggests that the data corresponding to the non-lips class is associated with higher complexity. Thus, it is intuitive to use a larger dictionary size for the non-lips class than the dictionary size for the lips class. In this manner, we not only control the data representation for each class but also consider the relative complexity differences between class data.

Next, we report the classification results for lips detection. Using the same dictionary size of 1000 for the lips and the non-lips class, the Dice score for test data was found out to be 35.28%. Using the information from PCA, we set the dictionary size for the non-lips class twice the dictionary size for the lips class. It was observed that the Dice score for the classification of test data was increased to 48.85% for the dictionary size of 1000 for the non-lips class and 500 for the lips class. This suggests that the information suggested by PCA about the complexity of the class data leads to improved classification accuracy, when we adapt the dictionary size for each class as hinted by PCA.

(a) Person 1 - Pose 1 (b) Person 1 - Pose 2 (c) Person 1 - Pose 3

(d) Person 2 - Pose 1 (e) Person 2 - Pose 2 (f) Person 2 - Pose 3

(g) Person 3 - Pose 1 (h) Person 3 - Pose 2 (i) Person 3 - Pose 3

Figure 5.2: Example of images in PUT Face data set. We selected 3 poses from the available 100 poses for three randomly selected persons in the data set. Each row in this figure shows three different selected poses for the same person.

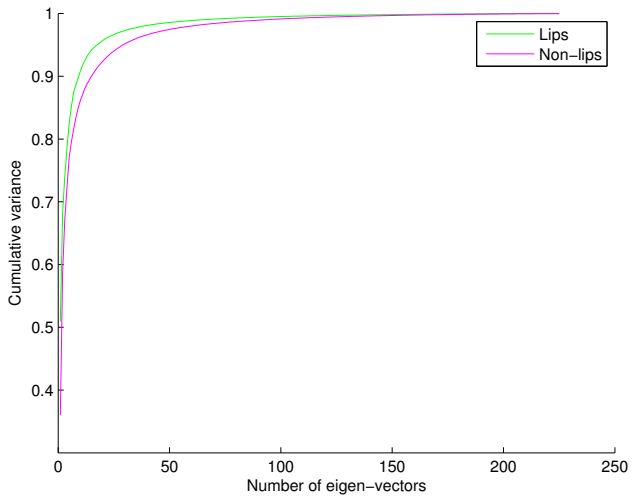


Figure 5.3: Principal componant analysis of the training data in the PUT Face database

It is however observed that the optimal Dice score was obtained for the dictionary size of 1000 for the non-lips class and 100 for the lips class. The value of the optimal Dice score was recorded as 73.03%. This suggests that PCA gives a hint about which class should have higher dictionary size but does not guarantee the optimal classification. One of the possible reason behind the failure of PCA to suggest the optimal dictionary size could be nonlinearity associated with the class data, where PCA might be inadequate in analyzing the data [Palŭs 1992].

The PCA of the USPS training data set in the similar manner is described in Table 5.2 and Figure 5.4. It is quite clear that the number of eigenvectors required to represent the data with 95%, 98% and 99% of cumulative variance for the digit 1 are very less as compared to all other classes. This indicates less complexity of the digit 1 as compared to rest of the digits. Therefore, PCA suggests the use of smaller dictionary size for digit 1 as compared to all other digits. The experimental results confirm the improvement in the classification accuracy by adapting the dictionary size according to PCA of the training data. The error rate using the same dictionary size of 255 for all classes is 3.44%, whereas the dictionary size of 255 for the digits except 1 and lower dictionary size, for example 100, only for the digit 1 reduces the error rate to 3.34%.

It is worth to mention here that the PCA can not be used to exactly estimate the ratio of dictionary size for all classes. One of the main reasons behind this is that PCA gives the best principal components, which are or-

	95%	98%	99%
0	69	116	150
1	20	37	53
2	103	148	178
3	97	144	175
4	89	132	161
5	92	140	172
6	72	110	137
7	64	102	130
8	94	134	159
9	69	105	129

Table 5.2: Principal component analysis of the training data for each digit class. For each class mentioned in a row, an entry in the table denotes the number of eigenvectors required to attain the percentage of total variance indicated in each column.

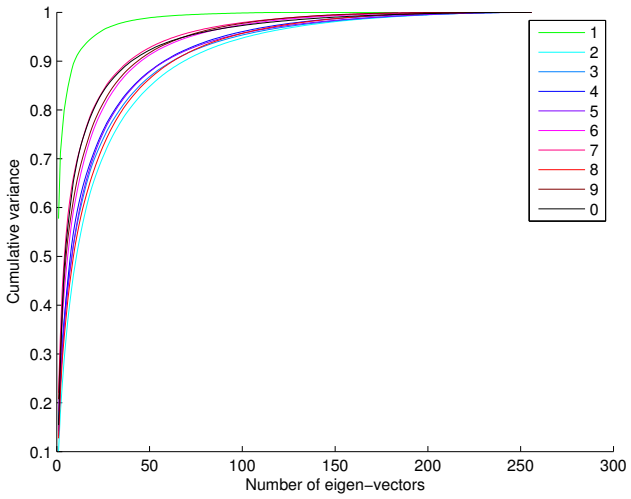


Figure 5.4: Principal Componant Analysis of the Training USPS Data Set

thogonal to each other and contain maximum variance within the given data. In the dictionary learning formulation, however, the columns of the dictionaries or the basis functions are redundant and they are not orthogonal to each-other as in PCA. We demonstrated the effectiveness of PCA in suggesting the differences in the complexity of the class data, as indicated by the relative number of eigenvectors required to reach specific level of cumulative variance. Therefore, the use of larger dictionary size is suggested for the class

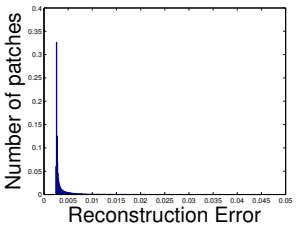
with higher complexity. It can be observed from the above experiments that even though PCA did not give the exact ratio of the dictionary size, the use of larger dictionary size for more complex class suggested by PCA resulted in improved classification.

5.2.2.2 Dictionary Size Selection using Histogram based Measures

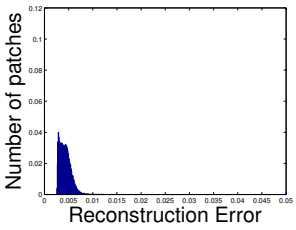
As described in section 5.2.1.2, we compute the histograms of the reconstruction errors of the lips and the non-lips class data using dictionaries for these classes, and calculate the Jeffrey divergence measure for comparing various histograms for the selection of optimal dictionary sizes. We performed the experiments on the training data by keeping the dictionary size of non-lips class constant, for example 1000, whereas the dictionary size for the lips class is varied from 1 to 1000 and the optimal dictionary size is selected by calculating histograms and the Jeffrey divergence measures, as described next.

For a fixed dictionary size of 1000 for the non-lips class, we calculated the histograms using the reconstruction errors of the training data of the non-lips and the lips class, as shown in Figure 5.5 (a) and (b), respectively. Throughout this experiment, the dictionary size for the non-lips class is kept constant and these histograms are to be compared against the histograms obtained using various dictionary sizes for the lips class. For several dictionary sizes for the lips class, the histograms obtained using reconstruction errors of the lips and the non-lips class using the class specific dictionaries for the lips class are shown in Figure 5.5 (c) - (n). The histograms on the left, (c), (e), ..., (m), indicate the representation ability of the dictionary for the lips class data. These histograms are compared against the histogram (a), which indicates the representation power of the non-lips dictionary for the non-lips class data. The Jeffrey divergence measure for this comparison, as denoted by $d_{J1}(H_{1,1}, H_{2,2})$ in section 5.2.1.2, is shown in red curve in Figure 5.6 (a). This term indicates the relative representation abilities of the dictionaries for the lips and the non-lips class data using the dictionaries for the corresponding classes. It can be observed that the value of Jeffrey divergence decreases as the dictionary size for the lips class is increased from 1 to 1000.

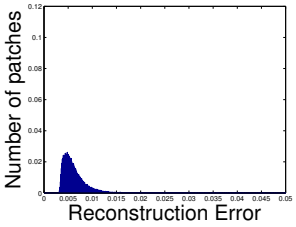
It is important that the dictionaries learned for each class are representative of their own class data, but are simultaneously not representative of the opposite class data. Figure 5.5 (b) shows the histogram obtained using the reconstruction error for the lips class and the dictionary for the non-lips class. Similarly, the histograms on the right, (d), (f), ..., (n) are obtained using the reconstruction error for the non-lips class data and the dictionaries of various sizes for the lips class. The comparison of histogram (b) with each of the histograms on the right, (d), (f), ..., (n), indicate how poorly the dictionar-



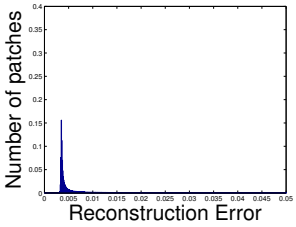
(a) Non-lips patches on the non-lips dictionary of size 1000



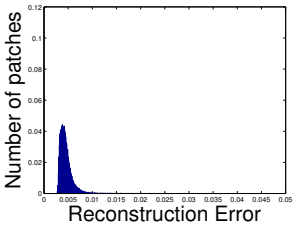
(b) Lips patches on the non-lips dictionary of size 1000



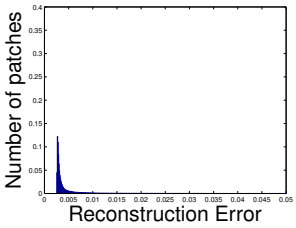
(c) Lips patches on the lips dictionary of size 10



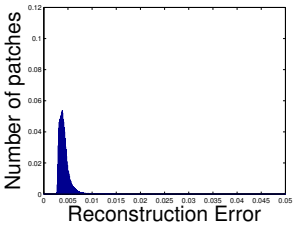
(d) Non-lips patches on the lips dictionary of size 10



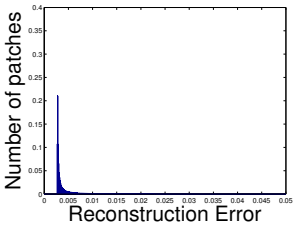
(e) Lips patches on the lips dictionary of size 50



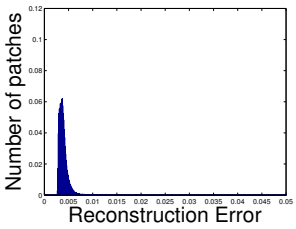
(f) Non-lips patches on the lips dictionary of size 50



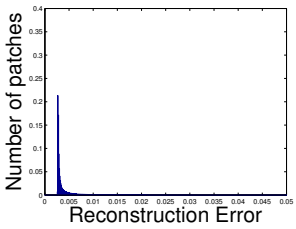
(g) Lips patches on the lips dictionary of size 100



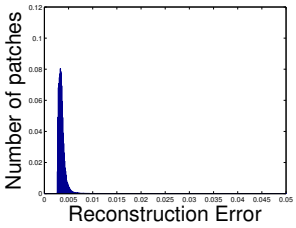
(h) Non-lips patches on the lips dictionary of size 100



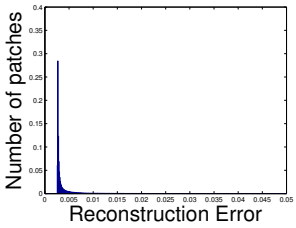
(i) Lips patches on the lips dictionary of size 200



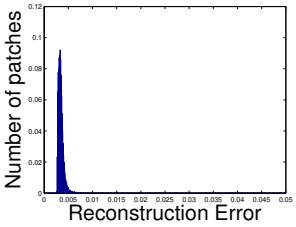
(j) Non-lips patches on the lips dictionary of size 200



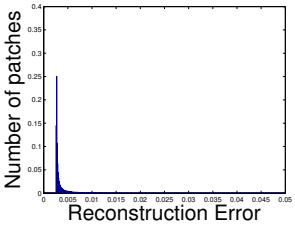
(k) Lips patches on the lips dictionary of size 500



(l) Non-lips patches on the lips dictionary of size 500



(m) Lips patches on the lips dictionary of size 1000



(n) Non-lips patches on the lips dictionary of size 1000

Figure 5.5: Histograms obtained using the reconstruction errors for the lips and the non-lips class, using the class specific dictionaries for the lips and the non-lips classes.

ies represent the data for the opposite class. The Jeffrey divergence measure for this comparison, as denoted by $d_{J2}(H_{1,2}, H_{2,1})$, is shown in blue curve in Figure 5.6 (a).

Figure 5.6 (b) shows the sum of squares of the Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ and $d_{J2}(H_{1,2}, H_{2,1})$. It is found that the minimum value is obtained for the dictionary size of 1000 for the non-lips class and 200 for the lips class. This is the optimal dictionary size for both the classes, which lead to the best representation and discrimination ability, owing to variability differences between the class data. We calculated the Dice scores for each of these dictionary sizes for the lips class, as shown in Figure 5.6 (c). It can be seen that the best classification is achieved at the dictionary size of 200 for the lips class. This confirms that the selection of dictionary size using histogram based method results in better classification.

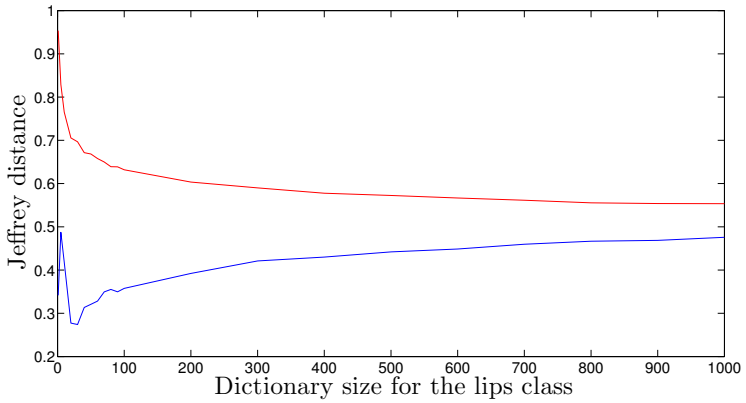
We performed the experiment above by selecting the face image from one of the training images. For the optimal dictionary size suggested by the histogram based measure, we achieved the best classification, as found experimentally. However, it was found that this method did not always give the exact dictionary size for the lips class for better classification, when experimented on the test images.

5.2.2.3 Dictionary Size Selection using Empirical Method

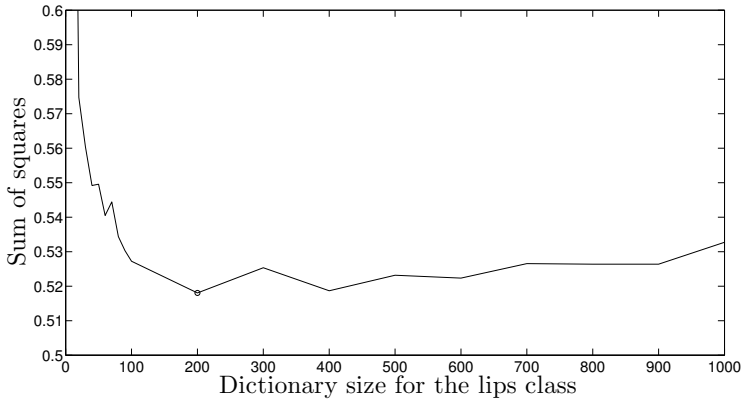
For this section, we use the data set divided into the training and test set by following a random 70%-30% split, as mentioned earlier. The dictionaries of various sizes are learned for the lips and the non-lips class, using the training data and the optimal dictionary size is selected by performing classification on the training data. These class specific dictionaries are then used to classify the test images.

First, we performed classification on the training data. The dictionary size for the non-lips class is kept fixed as 1000 and the dictionaries of size from 1 to 1000 are learned for the lips class. Figure 5.7 shows the variation of Dice score for different values of the dictionary size for the lips class while keeping the dictionary size for the non-lips class constant. The average Dice score for 210 training images using the same dictionary size is 40.32%. This value increases to 68.17% by using different dictionary sizes, 1000 for the non-lips class and 200 for the lips class. This experiment confirms that the dictionary size plays a major role in pattern classification. The optimal dictionary sizes as experimented on the training data, for the best classification, are 1000 for the non-lips class and 200 for the lips class.

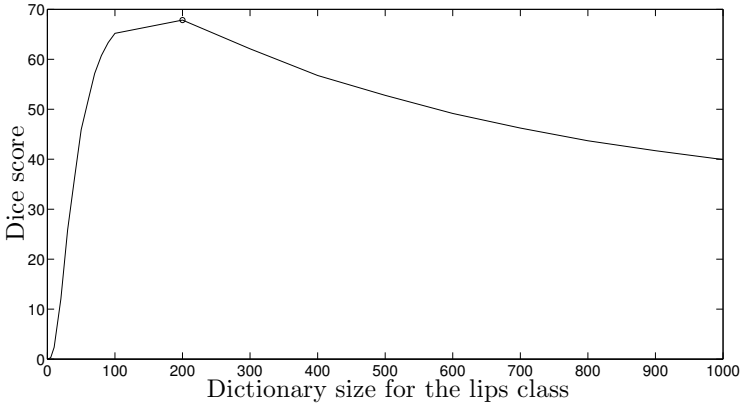
Next, we obtained the classification in similar manner for the test data. The dictionary size for the non-lips class is fixed to 1000 and the size of the



(a) Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ in red and $d_{J2}(H_{1,2}, H_{2,1})$ in blue, for the comparison of histograms



(b) Sum of squares of the Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ and $d_{J2}(H_{1,2}, H_{2,1})$. The minimum value is achieved at the dictionary size of 200 for the lips class, as indicated by the circled point on the curve.



(c) Dice scores for the blockwise classification of lips. The best classification is obtained at the dictionary size of 200 for the lips class, as indicated by the circled point on the curve.

Figure 5.6: The selection of dictionary size of the lips class using histogram based measures. The dictionary size for the non-lips class is kept constant as 1000 and the dictionary size for the lips class is caried from 1 to 1000. The optimal dictionary size for the lips class is chosen as 200, as indicated in Fig (b), which coincides with the best classification result obtained using this dictionary size, as shown in Fig (c)

dictionaries for the lips class is varied from 1 to 1000. We obtained the best classification for 90 test images, with Dice score of 73.03%, using dictionary size of 100 for the lips class. The Dice score with dictionary size of 200 for the lips class is very close to the best Dice score and is recorded as 69.39%. This suggests the success of the dictionary size selection using empirical method. The classification results for various dictionary size for the lips class are shown in Figure 5.8. It can also be observed that the average Dice score using the same dictionary size is 35.28% and is far worse than the best Dice score obtained using optimal dictionary sizes of 200 for the lips and 1000 for the non-lips classes suggested using classification on training data, and the dictionary sizes of 100 for the lips and 1000 for the non-lips classes found using the classification on test data.

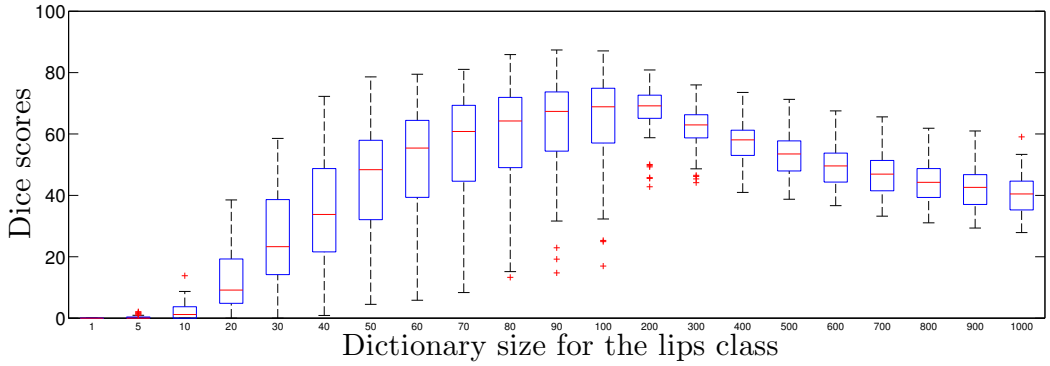


Figure 5.7: Dice scores for lips detection on training data, using SDL with a fixed dictionary size of 1000 for the non-lips class and the dictionary sizes of 1 to 1000 for the lips class.

5.3 Role of Dictionary Size in Discriminative Dictionary Learning

As discussed in the introduction of this chapter, the discriminative dictionary learning methods have been proposed with the objective of improving the classification accuracy. Several methods achieve this by introducing additional terms in the objective function of the dictionary learning formulation, so that the dictionaries learned from the given set of training data are reconstructive as well as discriminative. While these methods achieve better classification results in applications such as texture recognition, face recognition etc, it will be interesting to study if these methods achieve improvement by using an additional energy term in the dictionary learning formulation or by adapting

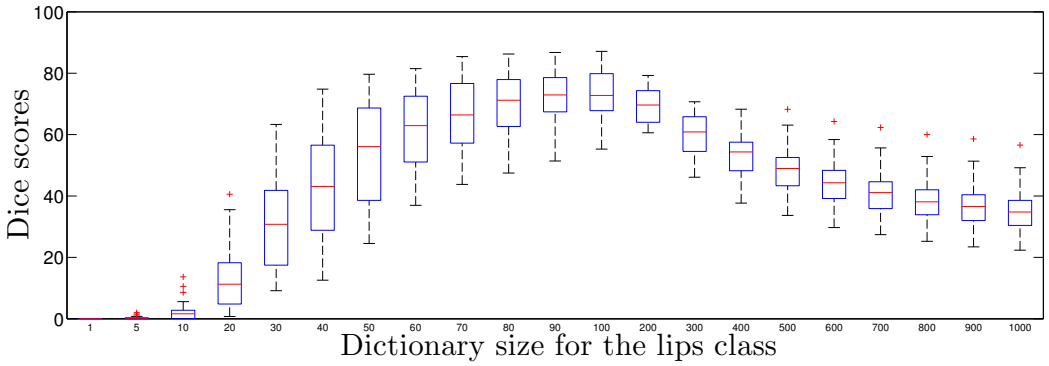


Figure 5.8: Dice scores for lips detection on test data, using SDL with a fixed dictionary size of 1000 for the non-lips class and the dictionary sizes of 1 to 1000 for the lips class.

the respective sizes of the dictionary for each class. This section is addressing this issue, which has not been studied so far, as per our knowledge.

We consider one of the most popular discriminative dictionary learning methods called Fisher Discrimination Dictionary Learning and investigate the role of dictionary size in the above mentioned application of lips detection in face images. The results of classification are compared with the standard dictionary learning method, using same and different dictionary sizes for each class.

5.3.1 Dictionary Learning Methods

In this section, we briefly recall the dictionary learning formulations, whereas the following sections describe the classification strategies employed using each of these methods. Finally, the results of lips detection application are discussed.

5.3.1.1 Standard Dictionary Learning (SDL)

For a set of signals $\{\mathbf{x}_i\}_{i=1,..,m}$, the dictionary learning problem is to find D such that each signal can be represented by sparse linear combination of its atoms. This can be stated as the following optimization problem

$$\min_{D,\{\alpha_i\}_{i=1,..,m}} \sum_{i=1}^{m} \|\mathbf{x}_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \tag{5.9}$$

The optimization is carried out as two step process involving the sparse coding step with fixed D and the dictionary update step with fixed α.

5.3.1.2 Fisher Discrimination Dictionary Learning (FDDL)

This method [Yang 2011] incorporates the Fisher discrimination criterion into the dictionary learning framework. The discriminative dictionary is obtained by solving

$$\min_{D,\alpha} R(X, D, \alpha) + \lambda_1 \|\alpha\|_1 + \lambda_2 f(\alpha) \tag{5.10}$$

where $R(X, D, \alpha)$ is a discriminative fidelity term, which ensures that each sub-dictionary for the corresponding class has a good representation power for the data from the same class, but has poor representation power for the other class data. The second term introduces a sparsity constraint and the last term $f(\alpha)$ is a discriminative coefficient term which uses the Fisher Criterion to minimize the within-class scatter and maximize the between-class scatter of sparse coefficients α. For more details, we refer the reader to the paper [Yang 2011].

5.3.2 Introduction to Method

The dictionaries learned from the training data are used for classification, using different classification methods described below.

5.3.2.1 Classification using Standard Dictionary Learning with Same Size (SDL-S)

Given the training data $X_i, i = 1, 2$ for 2 classes, we learn the dictionaries D_1 and D_2 of same size for the lips and the non-lips class, respectively, as described in Section 5.1.1.2. For a given test signal $\mathbf{y}$, the sparse coefficients α_c are calculated for each class $c = 1, 2$ using dictionaries D_1 and D_2, as mentioned in Section 5.1.1.3. Finally, the test patch is assigned to the class with a minimum reconstruction error, as described in Section 5.1.1.4.

5.3.2.2 Classification using Standard Dictionary Learning with Different Size (SDL-D)

In this method, we consider the variability differences between the lips and the non-lips class data, and allow larger dictionary size for the non-lips class.

5.3.2.3 Classification using Fisher Discrimination Dictionary Learning with Same Size (FDDL-S)

Using Eq. 4.15, we obtain a structured dictionary $D = [D_{F1}, D_{F2}]$ using FDDL, where D_{F1} and D_{F2} are the class-specified sub-dictionaries for the lips

Same Dictionary Size (200-200)

	PPV / Dice
SDL-S	22.9 / 36.9
FDDL-S	19.3 / 32.1

Different Dictionary Size (200-60)

	PPV / Dice
SDL-D	51.5 / 63.8
FDDL-D	**65.3** / **63.9**

Table 5.3: Results of lips detection for one test image. The table on the left indicates the Positive Predictive Value (PPV) and the Dice score, using two dictionaries of 200 atoms each for the lips and the non-lips classes, for SDL and FDDL. The table on the right indicates the PPV and Dice score for the adapted dictionary sizes: 200 atoms for the non-lips class and 60 for the lips class.

and the non-lips class, respectively. We then calculate the sparse coefficients and the metric for final classification as mentioned in [Yang 2011]. Following their recommendation, we selected the local classifier since the number of training samples for each class are large.

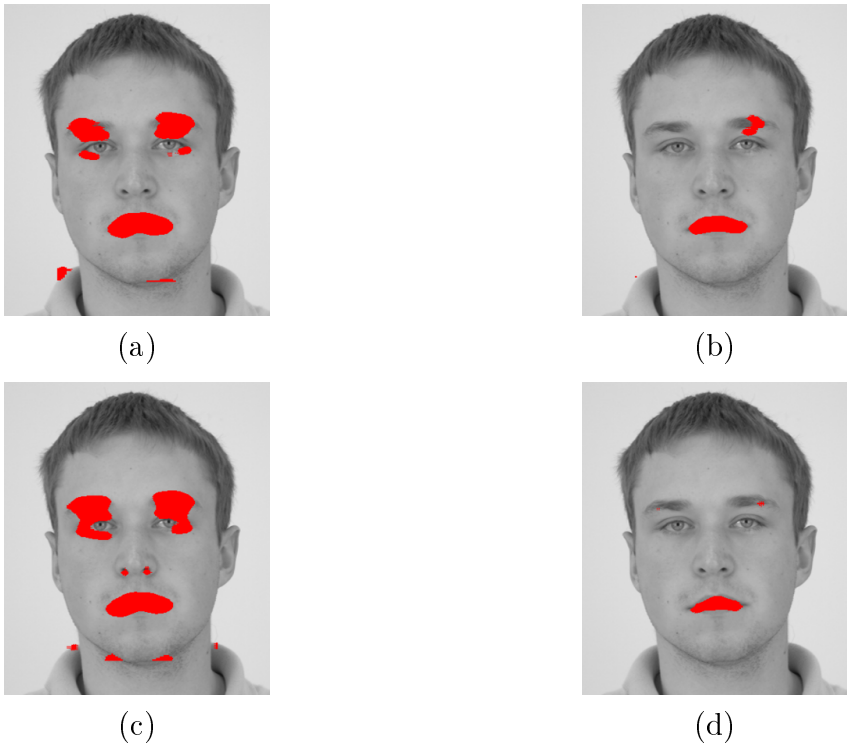
5.3.2.4 Classification using Fisher Discrimination Dictionary Learning with Different Size (FDDL-D)

To study the effect of dictionary size in this discriminative dictionary learning technique, we learn the dictionaries of different size using FDDL and obtain the classification as mentioned above.

5.3.3 Experiments and Results

The dictionaries obtained using the standard dictionary learning method are used as initialization dictionaries for FDDL algorithm. The sparsity parameter of $\lambda = 0.95$ was found to be the optimal choice for SDL, whereas the values of parameters chosen for FDDL were $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\gamma_1 = 0.1$ and $w = 0.1$. These values were chosen empirically after intensive testing to achieve the best results.

For using FDDL, the time-complexity was an important issue. The algorithm required long execution time when the training data size increased. We, therefore, sub-sampled the training data for the non-lips class by randomly selecting 1000 patches for each of the three poses for 70 persons and obtained the classification results for a randomly selected test image, as shown in Table 5.3. First, the classification is obtained using SDL method with same dictionary size of 200 for the lips and the non-lips class. This is followed by classification using SDL with different dictionary sizes. The dictionary size of 200 is kept constant for the non-lips class and the dictionary size for the lips



(a) (b) (c) (d)

Figure 5.9: Classification results for a randomly selected test image using: (a) SDL method with the same dictionary size of 200, (b) SDL method with different dictionary sizes: 200 for the non-lips class and 60 for the lips class, and (c) FDDL method with the same dictionary size of 200, and (d) FDDL method with different dictionary sizes: 200 for the non-lips class and 60 for the lips class. Lips detection is shown in red.

class is varied from 1 to 200. The best classification result is obtained for the dictionary size of 60 for the lips class. These dictionaries were then used for FDDL initialization. The dictionaries are updated using FDDL algorithm, for both the same and different dictionary size experiments, and the classification is then achieved as described in the section above.

The results of classification are indicated in Table 5.3. It can be observed that, similar to SDL, there is a vast improvement in the classification when we use FDDL with different dictionary size, when compared with the classification using the corresponding method with same dictionary size. This shows that the use of different dictionary size also results in improving discrimination between class data even in the case of discriminative DL technique. On the other hand, it can also be seen that FDDL with different dictionary sizes produces better result than SDL with different dictionary sizes. However, this is not true for the same dictionary size experiments. FDDL with same

dictionary size does not result in better classification than SDL with same dictionary size. This indicates that the discrimination introduced by FDDL does not alone guarantee improved performance, but the dictionary size plays a major role in discrimination between class data and hence achieves better classification.

Figure 5.9 shows the lips detection images using methods described above. The methods which employ the different dictionary sizes achieve better classification results as compared to the corresponding methods with the same dictionary size. In addition, FDDL with different dictionary sizes outperforms SDL with different dictionary sizes.

5.4 Conclusion

The standard and discriminative dictionary learning techniques have shown promising results in computer vision and pattern classification. We discovered that the major improvement in pattern classification can be achieved by adapting the dictionary size for each class, in the case of both the standard and discriminative dictionary learning methods. We firmly believe that the dictionary size is not just one parameter among others, especially for the classification purpose where one compares the representation power of several dictionaries. To illustrate the generic nature of this assertion, we validated the proposition of using different dictionary sizes based on complexity of the class data in a computer vision application such as lips detection in face images. In the next chapter, we investigate the performance of the dictionary learning methods in more complex application such as medical imaging.

Chapter 6

Classification of Multiple Sclerosis Lesions

Contents

Having investigated the role of dictionary size in pattern recognition with an example of lips detection in face images in the previous chapter, we present in this chapter the dictionary learning based classification method in a more complex medical imaging application. We consider a clinically relevant problem of the classification of multiple sclerosis lesions using multi-channel magnetic resonance images and study the effect of dictionary size in the classification of these pathological patterns in the medical images. We further describe methods to select the dictionary size for an optimal classification. The role of dictionary size in the discriminative dictionary techniques such as Fisher Discrimination Dictionary Learning (FDDL) is finally presented.

6.1 Multiple Sclerosis

Multiple sclerosis (MS) is a chronic, inflammatory, demyelinating disease of the central nervous system (CNS). The causes of the disease are not yet fully known. It is however known that this disorder of CNS damages the protective insulation, known as myelin, surrounding the nerve fibers called axons. In some cases, the nerves within the CNS and entire remaining structures are damaged as well. This breakdown of myelin sheath is known as demyelinization, which impairs the functionality of axons to communicate nerve impulses between neurons, as shown in Figure 6.1. The name multiple sclerosis is derived from multifocal hardened tissues known as plaques or lesions resulting from this demyelinization.

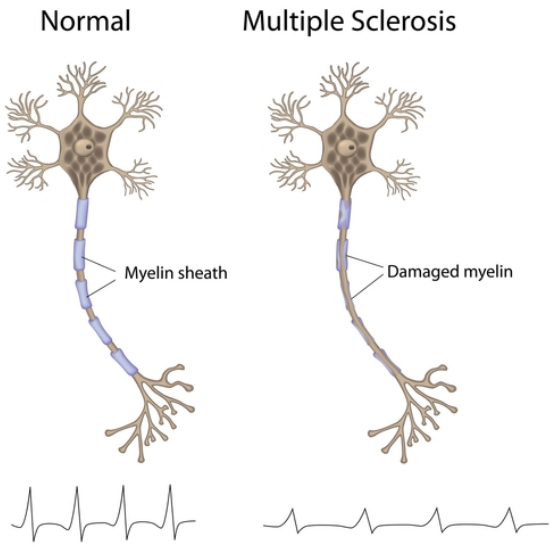


Figure 6.1: Demyelinization in Multiple Sclerosis. A healthy neuron is shown on the left and the diseased neuron with damaged myelin is shown on the right. [Espinosa 2014]

Multiple Sclerosis is believed to be an autoimmune disease, in which the immune system of the body itself attacks the body's own cells, causing inflammation in the CNS that destroys the myelin and the axons. This might result in symptoms as mild as numbness in the limbs to as severe as paralysis. Other symptoms include painful sensations, dizziness, muscle weakness, poor balance, slurred speech, fatigue, mood changes, loss of vision and susceptibility to depression. These symptoms can be temporary or permanent and can appear in any combination with different levels of severity. The nature and severity of the symptoms depend on location of the nerves where demyelinization has occurred and the intensity of the inflammation. The symptoms also

vary for each person, making it difficult for the doctors to determine the type and treatment plan for individual patient.

Multiple sclerosis is more common in North America and Europe and is more prevalent in young adult population. Approximately 400,000 people have been diagnosed with MS in the United states alone, with 200 new reported cases each week, the number of patients affected by MS worldwide are one million [Courtney 2006]. MS patients with first symptoms are diagnosed between the ages of 15 and 50. It is also observed that women are three times more susceptible to MS than men. The disease is prevalent among people raised in colder climates and although genetic factors make certain people susceptible to the disease, there is no scientific evidence that MS is inherited.

There are four disease courses in MS:

1. Clinically Isolated Syndrome (CIS): This is the first episode of neurological symptoms suggestive of MS, lasting at least 24 hours. The patient going through this episode may or may not lead to the development of MS.

2. Relapsing-Remitting MS (RRMS): This is the most common course of MS, with around 85% initial diagnosis. It is associated with attacks called relapses during which the old symptoms flare up or new symptoms are developed. Relapses are followed by a recovery time in weeks or months, called remission, during which some symptoms might disappear or some symptoms might continue to become permanent.

3. Secondary-Progressive MS (SPMS): In this type, symptoms steadily worsen over time, with or without relapses or remissions. The patients diagnosed with RRMS transition to this type after 10 to 20 years.

4. Primary-Progressive MS (PPMS): This type of MS is not very common and it occurs in about 10% of people with MS. It is characterized by gradual progressive worsening of symptoms from the beginning with little or no recovery.

Currently, there is no cure of MS, but the treatments which deal with different aspects of the disease are available. These include medicines that reduce the duration or shorten the severity of relapses, disease modifying agents that decrease the number of relapses, physiotherapy and medication to relieve the symptoms associated with MS and rehabilitation which consists of a therapy program to achieve and preserve the optimum physiological state. [Roberts 2006]

6.1.1 Magnetic Resonance Imaging for Multiple Sclerosis

Magnetic Resonance Imaging (MRI) is one of the most important modalities of medical imaging. It is capable of producing excellent contrast between tissues and allows to acquire multiple images of the same tissues with different contrasts with the help of different acquisition parameters and protocols. MRI is capable of providing high spatial resolution images, of the order of $1 \times 1 \times 1\ mm^3$ voxel size, and is an excellent imaging technique for studying the brain.

MRI holds the capability of detecting abnormalities in 95% of the patients with MS. It is the best paraclinical method for the diagnosis of MS, assessment of disease progression and treatment efficacy [Grossman 1998, Miller 2004]. The first MR images of MS were acquired in hospitals in 1980s and since then, MR has been used as a routine clinical examination in MS. The MR images are acquired every 3 months to 2 years for the detection of MS lesions, observe the status of the disease and to examine how well medications are working.

MRI achieves a great tissue contrast enabling the distinction between brain tissues; namely gray matter (GM), white matter (WM), and the cerebrospinal fluid (CSF). MS lesions most commonly occur in the white matter of the brain. Brain MR images highlight MS lesions in different intensity patterns depending on the MR modality used for the acquisition. Various MR modalities used for MS detection are T1-weighted, T2-weighted, T2-FLAIR, proton density, gadolinium-enhanced T1-weighted and Diffusion weighted imaging. For example, Figure 6.2 shows FLAIR, T1-weighted MPRAGE, T2-weighted and Proton Density (PD)-weighted MR images for a MS patient. Different MR sequences used for the diagnosis of MS in the clinical practice are described below.

1. In T1-w MR images, White Matter (WM) appears as the brightest tissue, when compared to Gray Matter (GM) as darker and the Cerebrospinal Fluid (CSF) as the darkest tissue. Active MS lesions appear as hypointense signal, while necrotic lesions, also known as black holes, are hypointense signals in T1-w sequence and are indicative of permanent nerve damage.

   T1-w images are also widely used for the diagnosis of MS, with an administration of contrast agent such as Gadolinium. When injected through a person's vein during a MRI scan, the flow of the Gadolinium based contrast agent into the brain or spinal cord is blocked by the blood-brain barrier. However, an inflammatory process in a lesion disrupts the blood-brain barrier and allows the passage of Gadolinium into the

brain or spinal cord. This results in the shortening of the longitudinal relaxation rate of the tissue, which subsequently results in a signal enhancement as seen on T1-w images.

2. Lesions appear hyperintense in both T2-w and PD-w images. While WM appears darkest in both images, CSF is brightest and GM is intermediate grey in T2-w images. One of the major drawback of using only T2-w images for MS lesion diagnosis is that the demyelination, inflammation, axonal loss, edema or gliosis lead to a hyperintense signal on T2-w images. Each of these pathologies are reflective of different stages of disease and are associated with different prognosis.

   In T2-w images, the lesions and CSF both appear with a high image intensities. This makes it difficult to segment lesions near the CSF-filled ventricles. Proton density images have a reduced signal intensity for CSF as compared to T2-w images and these images could be acquired together with T2-w images in the same sequence.

3. In FLAIR, CSF signal is suppressed so that it appears darkest, while WM appears intermediate gray and GM appears brighter than WM and CSF. MS lesions appear as bright signal. FLAIR images are better choice for detection MS lesions present on the boundary of the ventricles. The only disadvantage with FLAIR image is the requirement of higher acquisition time.

4. Diffusion-weighted imaging (DWI) MR scans provide information about water diffusion in tissues. There is increased amount of water diffusion in the regions in the brain which are affected by MS. This causes signal changes in DWI images in the presence of MS lesions, allowing the examination of the type, appearance and location of MS lesions [Goldberg-Zimring 2005].

MRI is a non-invasive technique, which does not utilize ionizing radiation and has no side-effects. It is therefore best suited for the repeated examination of MS patients, which allows to study the progression of disease over the course of time and the effect of drugs on the evolution of the disease. Technological advances of MR in recent years have dramatically improved our understanding of MS.

6.1.2 Diagnostic Criteria for MS

Until the end of 20th century, two popular diagnostic criteria for the diagnosis of MS were Schumacher criteria, developed in 1965 and Poser criteria, pro-

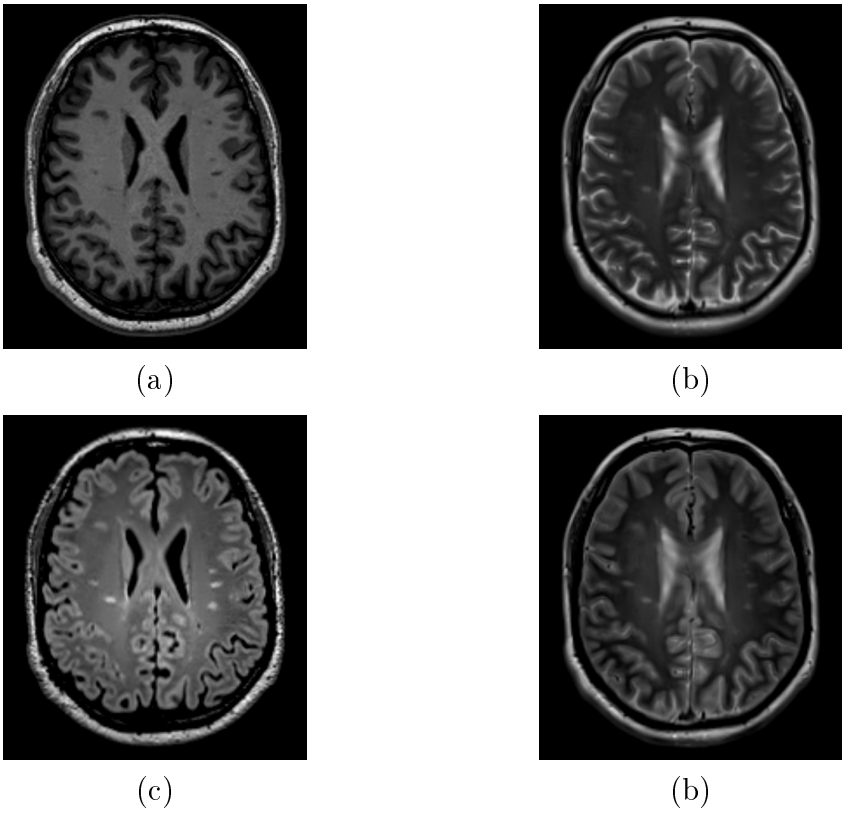


Figure 6.2: MR images of brain with MS lesions. (A) T1-w MPRAGE, (B) T2-w, (C) FLAIR and (D) PD-w, respectively.

posed in 1983. These criteria were purely clinical and they were used before MR imaging was proved to be a standard tool for the diagnosis of MS.

Conventional MR image sequences provide good sensitivity in the detection of MS lesions and quantitative assessment of lesion load. The quantitative parameters derived from these MR images have helped in understanding the natural history of the disease and monitoring of the disease progression for efficient treatments. McDonald criteria, originally published in 2001, uses the increased understanding of the natural history of MS as obtained from MR images and clinical progression for the diagnosis of MS. It was subsequently updated in 2005 and 2010 [Polman 2005, Polman 2011]. This criteria proposed the outcome of a diagnostic evolution as either "MS", "possible MS" or "not MS", instead of previously used terms such as "clinically definite" and "probable MS" in the earlier criteria [McDonald 2001]. It is based on two main components: The presence and spatial pattern of the lesions (dissemination in space) and the appearance or disappearance of the lesions (dissemination

in time). The latest version of the criteria allows the early diagnosis of MS with a high degree of specificity and sensitivity.

Several authors have proposed criteria that classify MR image findings such as lesion number, location and various other characteristics to indicate the possibility of MS [Barkhof 1997, Tintoré 2000]. Therefore, for the diagnosis of MS, the MR images are analyzed to find the number and spatial patterns of the lesions, appearance of new lesions and the total lesion load, which are key parameters in the current MS diagnostic setup.

6.1.3 MS Lesions Segmentation

Manual segmentation of MS lesions is a laborious and time consuming task, pertaining to the requirement of analyzing a large number of MR images. It demands for an expert neurologist or radiologist, and there exist inconsistencies in the manual delineation of lesions among experts. Low lesion contrast, irregularities in the common intensity and texture characteristics, unclear boundaries resulting from the partial volume effect and the changing tissue properties are the main causes of error and the intra- and inter-expert variability. Furthermore, there are additional challenges in MS lesions segmentation as the shape and location of the lesions within white matter varied across patients. These problems become more prevalent as the number of MR modalities used for the diagnosis increases. Analyzing several 3D MR volumes keeping in mind the contrast differences between tissues and the intensity characteristics of MS lesions in each MR modality adds more complexity in manually segmenting the MS lesions for large number of patients. Therefore, fully automated methods, which guarantee good accuracy and reproducibility, along with the reduced processing time, are required for the segmentation of MS lesions. Several automatic or semi-automatic MS lesion segmentation have been proposed over the last decades, with an objective of handling a large variety of MR data and which can provide results that correlate well with expert analysis. We provide a brief review of these methods as described next.

The manual segmentation images obtained from the experts are considered to be the silver standard since they provide the best in-vivo estimate available but they are not the perfect ground truth representations. Different modalities are examined by the experts for the selection of the lesion voxels and this complicated process may result in different experts reporting different results or the same expert reporting different results on the same MS patient on each different evaluation. Computerized methods provide benefits in analyzing the complex multiple MR modalities while effectively utilizing the information from multiple adjacent slices. This has resulted in learning accurate models

for segmenting MS lesions. These techniques utilize various methodologies from different streams of science and consist of comprehensive frameworks made of several steps, including pre- and post-processing.

The segmentation algorithm first pre-process MR images for the removal of noise, motion, partial volumes, anatomical variations and blurred edges, which may degrade the results of subsequent image analysis and pose additional challenges in MS lesions segmentation. These methods incorporate the following pre-processing steps: (1) Noise reduction: The noise induced by the acquisition process is first eliminated [Coupe 2008]. (2) Intensity inhomogeneity (IIH) correction: The inhomogeneity of the static or applied magnetic fields within the MR acquisition device causes intensity variations of the same tissue with respect to the locations of the tissues. IIH correction methods reduce these intensity inhomogeneities, which subsequently improves the segmentation [Vovk 2007]. (3) Intensity normalization: Some segmentation methods require uniform intensity patterns within the training data set, testing data set and longitudinal MS lesions image data. The intensity range of the target image is modified and mapped into a predefined intensity range [Nyul 2000, Karpate 2014]. (4) Registration: This step registers MR images into the same space, so that all images to be processed further are brought into the best possible spatial correspondence with respect to one another [Maintz 1998]. (5) Skull stripping: This is another important preprocessing step as the non-brain tissues have intensity similarities with brain structures and this may cause mis-classifications in some approaches. There exist several approaches for brain extraction, which allow the segmentation to be performed on the selected brain voxels [Smith 2002].

The MS lesions segmentation methods can be distinguished in terms of features each of these methods use. As discussed earlier, each MR modality such as T1-w, T2-w, PD-w and FLAIR represent healthy brain tissues and the lesions in different intensity patterns. The proposed approaches can be classified in terms of whether they use single channel or multi-channel MR images, which act as a set of features for their algorithm. Some approaches use T1-w images for tissue segmentation, because of a good contrast differences between tissues in T1-w images, and the initial tissue segmentation is then used to obtain the lesion segmentation. There exist approaches using single channel MR image such as FLAIR sequence for obtaining the segmentation [Khayati 2008, Weiss 2013, Abdullah 2011]. On the other hand, the use multi-channel MRI increases the intensity feature space and produces better segmentation as a result of better discrimination between brain tissues. Several approaches have been proposed that use more than one of T1-w, T2-w, PD-w and FLAIR images for the segmentation of MS lesions [Prastawa 2008, Akselrod-Ballin 2009].

These methods are based on semi-supervised, supervised or unsupervised approach and use different classification strategies to model lesions [Lladó 2012, Mortazavi 2012, García-Lorenzo 2013].

6.1.3.1 Supervised Approaches

The algorithms proposed in this category use the training data in the form of manual segmentation images to learn the characteristics of the lesions. The features extracted from the manually segmented image are fed to the classifier, which is trained to perform the segmentation of MS lesions. Training database needs to be chosen carefully in such approaches so that heterogeneity of MS lesions and the variability of MR acquisitions are taken into account.

Some classification approaches implement binary classifiers to classify the final output as lesion or not lesion, while other methods use multiple labels for each tissue and produce a probabilistic map, which can be processed to obtain the lesions segmentation. In most of the approaches, the features are extracted using the manual segmentation image and the classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) or Bayesian classifier are trained. Majority of the algorithms use multi-channel MR images.

KNN classification technique was used for the automatic classification of WM lesions using voxel intensities and the spatial information as the features. T1-w, Inversion Recovery (IR), PD, T2-w and FLAIR MR modalities were used [Anbeek 2004]. The advancement over this method using KNN as a classifier is proposed for simultaneous segmentation of lesions and brain tissues. The technique generates the probabilities of a voxel belonging to lesion or brain tissue, which are followed by the application of thresholds to obtain the final segmentation [Anbeek 2005].

The approaches for the MS lesions segmentation using SVM focus on the extraction of relevant image features and finding the optimal decision boundary so as to achieve a maximal separation between the classifying hyperplane and the samples on the margin called the support vectors. The non-linear data can be transformed into a different feature space, with the use of kernels such as radial basis function [Ferrari 2003]. The use of multiple modalities might not always lead to the improved performance using these approaches. It was shown that the use of FLAIR and T1-w images gives a similar performance at a lower cost, when compared with the results using FLAIR, T1-w, T2-w and PD images [Fiot 2013].

The intensity and geometric properties of lesion were considered for building ANN based framework for the lesion segmentation. Multi-sequence MRI data is used and the hyper-intense regions in image are identified using adap-

tive threshold algorithm applied on normalized images. The artifacts in this step are partially removed by considering morphological properties including area and shape. Finally, ANN is trained to segment the lesions. A back-propagation architecture with three layers 3-5-2 is considered, where the shape index, average intensity and the product of these two form three inputs and the outputs are two classes indicating lesion or non-lesion. The weights are then learned using supervised learning approach using the back-propagation algorithm [Goldberg-Zimring 1998].

Some supervised approaches use the MS probability atlas constructed from the expert annotated lesions in the training data, along with the image features such as neighbourhood voxel intensities, the derivatives of the voxel intensities and the histogram information which provides the low pass intensity information of a certain region. Principal Component Analysis (PCA), with a log-likelihood ratio, is then used to classify each voxel [Kroon 2008]. The benefit of such atlas-based approach is that it inherently uses the spatial information. Other methods, on the other hand, use real characteristics of the tissues and the lesions, but spatial information has to be incorporated as an additional step.

Several other approaches using different machine learning techniques for the lesions segmentation include Bayesian frameworks [Harmouche 2006], decision trees [Kamber 1995], logistic regression [Sweeney 2013], least squares probabilistic classifier [Karpate 2015] and deep learning [Brosch 2016]. Ensemble of classifiers, which combine several base learners to produce a strong classifier, are also used for the lesions segmentation. The approach using 3D features based on multichannel intensity, prior and context-rich information, and a spatial random decision forest classifier is one such example [Geremia 2010]. The other approach in this category uses intensity and contextual features along with an extended version of the outlier map, and a boosting classifier to achieve the MS lesions segmentation [Cabezas 2013].

6.1.3.2 Unsupervised Approaches

Unsupervised methods do not require labelled training data in order to perform the segmentation. For these methods, although the complex process of manual segmentation can be avoided, the translation of expert knowledge and unsupervised classification methods to first segment brain tissues to help lesion segmentation or directly use the lesion properties to segment MS lesions is a challenge.

Many intensity based unsupervised approaches were proposed to classify the healthy brain tissues into three classes: WM, GM and CSF. A fuzzy C-mean and a finite Gaussian mixture model with the expectation maximization

(EM) algorithms were used for this purpose [Wells 1996a, Pham 1999]. The lesions are detected by adding separate class for MS lesions [Souplet 2008] or treating lesions as outliers [Leemput 2001]. The later method uses multi-sequence information, removes MR field inhomogeneities and incorporates contextual information in the classification using a Markov random field. The advantage of this method is that it eliminates the modeling of lesions and this results in robust estimation in the presence of other tissues or artifacts.

Another approach combined two segmentation methods, the Mean Shift and a variant of the EM algorithm to segment MS lesions. The Mean Shift uses local information to generate number of regions in the images, which are merged using neighboring information. A variant of EM, using trimmed likelihood estimator, is employed to classify the regions obtained into normal appearing brain tissues (NABT) or lesions [García-Lorenzo 2008]. In another work, the maximum likelihood estimator is replaced by a robust likelihood estimator to avoid the outliers in the estimation. The segmentation is refined using both the Mahalanobis distance of intensity of WM voxels and prior information coming from clinical knowledge on lesion appearance across sequences. The algorithm is validated using 3D + t MR data to segment MS lesions over time [Aït-Ali 2005].

While most algorithms use only intensity information, several algorithms are also proposed which use the spatial information in order to improve the lesions segmentation [Leemput 2001, Khayati 2008]. In these approaches, Markov Random Field (MRF) is incorporated to include the local neighborhood in the estimation and the lesions are identified as outliers not correctly explained by the model.

6.2 Dataset and Preprocessing

The dataset chosen for the MS lesions segmentation approach consists of MR images of 13 patients acquired via 3T Siemens Verio (VB17) scanner. T1-w MPRAGE, T2-w, PD-w and FLAIR modalities were chosen for the analysis. The volume size for T1-w MPRAGE and FLAIR was $256\times256\times160$ and voxel size was $1\times1\times1$ mm^3. For T2-w and PD-w, the volume size was $256\times256\times44$ and voxel size was $1\times1\times3$ mm^3. Annotations of the lesions were carried out on T2-w volume by an expert neuroradiologist. These manual segmentation images are referred to as ground truth lesion masks.

The noise introduced during MR acquisition is removed using non-local means [Coupe 2008] and intensity inhomogenity (IIH) correction [Tustison 2010]. To ensure the spatial correspondence, the images are registered with respect to T1-w MPRAGE volume [Wells 1996b] and are processed further to extract

the intra-cranial mask [Smith 2002]. We limit our further analysis to this brain region.

6.3 MS Lesions Segmentation: 2-Class Method

Over the last few years, sparse representation has evolved as a model to represent an important variety of natural signals using a linear combination of a few atoms of an over-complete dictionary. Dictionary learning, a particular sparse signal model, aims at learning a non-parametric dictionary from the underlying data. The representation of data in such a manner has led to the use of sparse representations and dictionary learning in many image processing applications such as image restoration [Elad 2006c, Mairal 2008a], inpainting [Elad 2010], face recognition and texture classification [Peyré 2009, Wright 2009].

The ability of sparse representations to approximate high-dimensional images using a few representative signals in a low-dimensional subspace and the development of efficient sparse coding and dictionary learning techniques offer a great advantage in medical image analysis. Recent publications have demonstrated the effectiveness of sparse representation techniques in medical applications such as shape modelling [Zhang 2012a], constructing a structural brain network model [Chung 2011] and predicting cognitive data from medical images [Kandel 2013]. In addition, the dictionary learning framework has been used in deformable segmentation [Zhang 2012b], image fusion [Yu 2013], super-resolution analysis [Wang 2012], denoising [Rubinstein 2010b, Deka 2010], deconvolution of low-dose computed tomography perfusion [Fang 2013a, Fang 2013b] and low-dose blood-brain barrier permeability quantification [Fang 2014]. In each of these applications, the dictionaries are learned from the underlying data so that they are better suited for representation of the signal of interest. On the other hand, the discriminative dictionary learning approaches proposed for image segmentation focus on learning dictionaries which are representative as well as discriminative [Zhang 2010b, Tong 2013]. In this work, we propose a novel algorithm, for the classification of multiple sclerosis lesions, which incorporates discrimination in the dictionary learning framework by varying the size of the dictionaries according to the complexity of the underling data. Very few approaches proposed in the past have considered the effects of the dictionary size in image classification [Ramirez 2012, Gao 2014]. We investigate this in the particular case of classification in the medical imaging application.

In the past, Weiss *et al.* [Weiss 2013] proposed dictionary learning based MS lesion segmentation method by learning a single dictionary with the help of healthy brain tissue and MS lesions patches. The lesions are treated as

outliers and lead to a higher reconstruction error when decomposed using this dictionary. There are several shortcomings in this method. The method uses only FLAIR MR images for analysis of clinical data. However, MS lesions appear in different intensity patterns in various MR sequences, which include T1- (T1-w MPRAGE), T2- (T2-w) and Proton Density-weighted (PD-w). The complementary information in these MR images can further assist in classifying MS lesions. We, therefore, build our analysis using multi-channel MR data.

The former method also uses an unsupervised approach and it was observed that one of the crucial parameters used in this approach is the threshold on error map. This parameter drives the segmentation results and is not easy to tune. Furthermore, it could lead to worse segmentation results for small errors in the brain extraction procedure. We suggest a solution to this problem by proposing a fully automatic supervised classification method that eliminates this parameter. As outlined in many classification approaches using dictionary learning, we learn class specific dictionaries for the healthy brain tissues and the lesions that promote the sparse representation of the healthy and lesions patches, respectively. The lesions patches are well adapted to their own class dictionary, as opposed to the other. Thus, we can use the reconstruction error derived from the sparse decomposition of the test patch on to these dictionaries for obtaining the classification. Finally, the effect of the dictionary size for the healthy brain tissues and the MS lesions class in the classification of MS lesions is investigated.

6.3.1 Methodology

As shown in Figure 6.3, we first preprocess MR images for noise removal and then extract the image patches of predefined size using brain mask. These patches are normalized and are divided into the training and test sets for healthy brain tissue and the lesions classes, with the help of manual segmentation images. Using training signals, we derive different classification approaches by either learning single dictionary or two separate dictionaries for both the classes. Finally, for a given test patch, the reconstruction error based classification method is developed, followed by voxel-wise classification and the lesions detection. The following subsections briefly describe these steps.

6.3.1.1 Patch Extraction and Training Set

We divide the intracranial MR volume into several 3-D patches and flatten them into one dimensional concatenated vectors representing intensities of

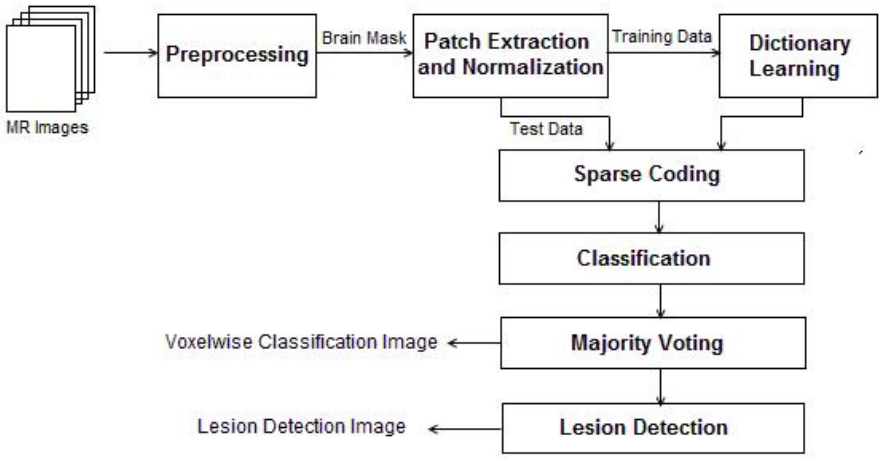


Figure 6.3: Flowchart of MS Lesions Detection using Dictionary Learning (2-Class Method)

T1-w MPRAGE, T2-w, PD and FLAIR images. Keeping the computational complexity of further analysis in mind, we extract a patch every M voxels in each direction. As described earlier, we develop supervised approach by labelling these patches as belonging to either healthy or the lesions class. If, in a patch, the number of voxels manually labelled as lesions exceeds a threshold $T_L = 6$ voxels, it is included in the lesions set, or in healthy set otherwise. For every subject, we obtain around 1.5×10^6 patches for healthy and 10^3 to 10^5 patches for the lesions class, depending on the lesion load for each patient. These patches are finally normalized to limit their individual norms below or equal to unity, as per constraint imposed by dictionary learning.

6.3.1.2 Dictionary Learning and Sparse Coding

Sparse representation of the data allows the decomposition of signal into linear combination of few basis elements in an overcomplete dictionary. Consider a signal $\mathbf{x} \in \mathbb{R}^N$ and an overcomplete dictionary $D \in \mathbb{R}^{N \times K}$. The sparse coding problem can be stated as $min_\alpha \|\alpha\|_0$ s.t. $\mathbf{x} = D\alpha$ or $\|\mathbf{x} - D\alpha\|_2^2 \leq \varepsilon$, where $\|\alpha\|_0$ is l_0 norm of the sparse coefficient vector $\alpha \in \mathbb{R}^K$ and ε is error in representation. Basis pursuit algorithm solves the convex approximation of the problem above by replacing l_0 norm with l_1 norm that also results in sparse solution [Chen 1998]. Thus, the sparse coding problem can be given by

$$\min_\alpha \|\mathbf{x} - D\alpha\|_2^2 + \lambda \|\alpha\|_1 , \tag{6.1}$$

where λ controls the trade-off between representation error and sparsity.

The fixed dictionaries like wavelets can be efficient if a background analytical model can be inferred. On the other hand, the dictionary learning from underlying data has produced exciting results with greater data adaptability and has replaced the use of generic models. For a set of signals $\{\mathbf{x}_i\}_{i=1,..,m}$, the dictionary learning problem is to find D such that each signal can be represented by sparse linear combination of its atoms. This can be stated as the following optimization problem

$$\min_{D,\{\alpha_i\}_{i=1,..,m}} \sum_{i=1}^{m} \|\mathbf{x}_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 . \tag{6.2}$$

The optimization is carried out as a two-step process involving the sparse coding step with fixed D and the dictionary update step with fixed α.

6.3.1.3 Patch Classification

We use three different strategies for dictionary learning and the classification of test patches. These methods are explained below.

(a) Single Dictionary (SD)

In the context of MS lesion classification, the simplest idea, similar to [Weiss 2013], could be to use a single dictionary learned from healthy and the lesions class patches. As the lesions are outliers with respect to the healthy brain intensities, the decomposition of lesion patch using this dictionary would result in higher representation error than that for the healthy tissue patch. For a given test patch, we calculate the sparse coefficients and reconstruction error, and assign it to the lesions class if this error is greater than chosen threshold. The threshold is selected by observing the histogram of the error map.

(b) 2-Class Specific Dictionaries - Same Size (2D-S)

Here, we learn class specific dictionaries D_1 and D_2 of same size for the healthy and the lesions classes, respectively. Given a test patch $\mathbf{x} \in \mathbb{R}^N$, the classification is performed in two steps: In the first step, sparse coefficients α_i are obtained using Eq (1) for each class $i = 1$ (Healthy) and 2 (Lesions). The test patch is then assigned to class c such that

$$c = \underset{i}{\operatorname{argmin}} \|\mathbf{x} - D_i\alpha_i\|_2^2 . \tag{6.3}$$

(c) 2-Class Specific Dictionaries - Different Size (2D-D)

The dictionaries learned using above mentioned approach does not take into account the data variability between two-classes. As demonstrated

in the previous part, the size of the dictionary plays a major role in the data representation. For the healthy class data with more variability and number of training samples than that for the lesions class, we allow larger dictionary size for healthy class data and study its effect on MS lesion classification.

6.3.1.4 Voxel-wise Classification and Lesion Detection

As already stated, there is some overlap between patches. However, to obtain voxel-wise classification, each voxel needs to be assigned to either of the classes. This is achieved using majority voting, in which, the voxel under consideration is classified as healthy or lesion, using majority votes of all patches which contain that voxel.

The voxelwise classification image is further processed to obtain the lesion based detection image. A lesion is said to be detected if $\frac{R_D \bigcap R_{GT}}{R_{GT}} \geq T_O$, where R_D and R_{GT} are respectively the candidate regions in the classification image and the ground truth, whereas T_O is the threshold indicating overlap between them as a fraction of ground truth lesion.

6.3.2 Results and Discussions

We implemented our method using MATLAB and Python. The packages ANIMA[1] and N3 ITK were used for denoising, registration and IIH correction, respectively [Coupe 2008, Wells 1996b, Tustison 2010]. We used the neuroimaging software Brain Extraction Tool (BET) for the brain extraction [Smith 2002]. For dictionary learning and sparse coding, we used SPArse Modeling Software (SPAMS) package [Mairal 2009b].

We performed the experiments on 13 subjects using Leave-One-Subject-Out-Cross-Validation. Different parameters have been tested for the methods. It was found that image patch of size $5 \times 5 \times 5$, with a patch every 2 voxels in each direction, was optimal with respect to the classification efficiency. For voxel-wise classification method, we then recorded the number of voxels that belong to True Positives (TP), False Negatives (FN), False Positives (FP) or True Negatives (TN) and the classification methods were finally validated by calculating sensitivity= $\frac{TP}{TP+FN}$ and Positive Predictive Value (PPV) = $\frac{TP}{TP+FP}$.

In the first method, we studied the classification by learning single dictionary with the help of both healthy brain tissue and the lesions patches. We chose the dictionary size of 5000 and the sparse penalty factor $\lambda = 0.85$ in the sparse coding step. The classification is then performed for various

[1] https://github.com/Inria-Visages/Anima-Public

threshold values on the histogram of error map, as explained previously. We then selected the threshold for which the best voxelwise classification results were obtained in terms of both sensitivity and PPV. It was observed that the method suffered with a very large number of false positive detections, as shown in Figure 6.4(a).

Next, we learned the class specific dictionaries for the healthy and the lesions classes, each. We used dictionary size of 5000 for the signal representation of each class. The optimal value of the sparsity parameter λ was found to be 0.95. The mean sensitivity and PPV obtained using this approach were 95.8% and 7.9%. This method performs better than the previous method but still contains many false positives. The primary reason behind this can be the difference in the data variability of each class signals. The healthy class patches have more variability in terms of representation of white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF), as compared to the variations in the representation of lesions. Therefore, following the observations made in Section 6.3.1.3(c), we adopted different dictionary sizes for representation of these classes. We used dictionary sizes of 5000 and 1000 respectively, for healthy and the lesions classes. Table 6.1 summarizes the results of the voxelwise classification for the three methods described above.

It can be seen that using class specific dictionaries with the same dictionary size improves both sensitivity and PPV, as compared to the first method. But PPV in the second method is still low, indicating that there are still large number of false positives, which can explain higher sensitivity. Using different dictionary size for each class, as implemented in the third method, drastically reduces the number of false positives, which can be seen by the significant increment in PPV, while keeping the sensitivity in the acceptable limit.

The mean PPV and sensitivity for lesions detection with class specific dictionaries of different size are shown in Table 6.2 for various overlap thresholds T_O. To be consistent with the threshold T_L incorporated in learning stage (Refer Section 6.3.1.1), we ignore very small lesions with volumes less than $T_L = 6$ voxels. It can be seen that we detect 64.8% of the lesions with the overlap threshold of 1% . Moreover, in 53.36% of the lesions detected, at least 40% of the voxels are correctly classified by the method.

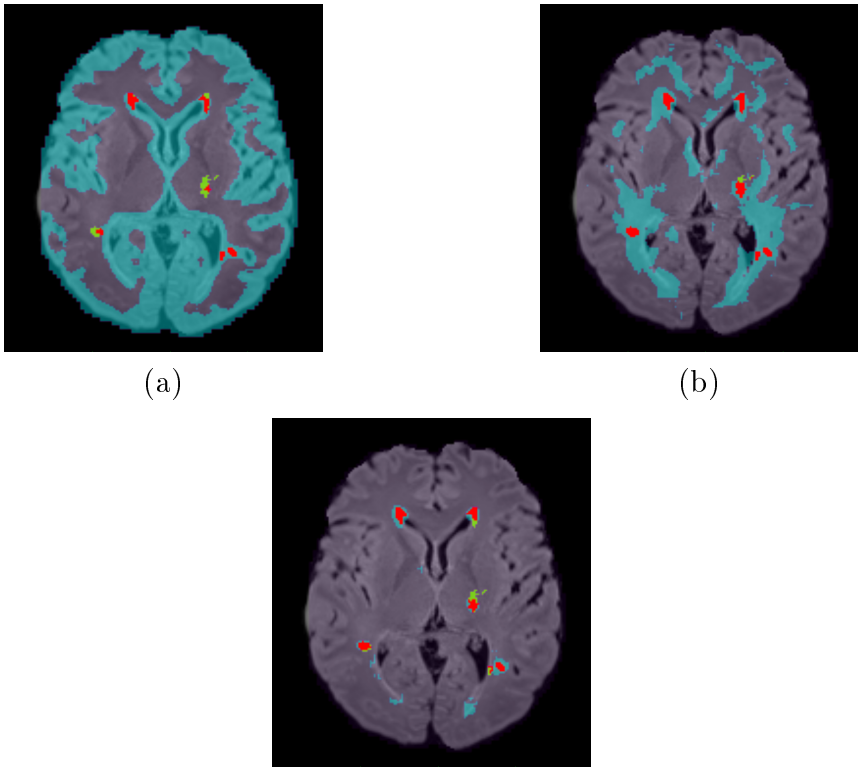
In Figure 2, we show the results for patient 8, for all the methods discussed above. The detection image is superimposed on FLAIR image. It can be observed that methods (a) and (b) have large number of false positives. We get the best classification results using class specific dictionaries with different dictionary sizes. But, in terms of voxelwise classification, there are still few false positives and true negatives around actual lesion. This does not pose a major problem for lesions detection as long as significant portion of the actual lesion is being classified correctly.

Pat. No.	(a) 1D			(b) 2D-S			(c) 2D-D		
	SEN	PPV	Dice	SEN	PPV	Dice	SEN	PPV	Dice
1	42	1	0.2	97	3	4.3	53	31	38.5
2	74	1	0.3	98	2	3.7	66	41	50.4
3	73	1	0.4	91	2	3	63	27	36.8
4	91	2	2.3	98	17	27.9	57	68	61.4
5	61	1	1.2	95	10	18	54	65	58.8
6	91	7	12.4	89	29	42.9	38	55	44.4
7	78	1	0.5	85	3	5.3	20	32	24.2
8	72	1	0.8	98	3	4.4	69	21	31.6
9	66	1	1.2	97	9	15.2	61	52	55.7
10	89	2	3.6	98	12	21.2	66	41	50.3
11	75	1	1.4	99	8	13.5	52	36	42.3
12	78	1	0.9	100	3	5.3	77	31	43.8
13	59	1	0.3	100	2	2.3	78	17	27
Mean	**73**	**1.6**	**2**	**95.8**	**7.9**	**12.8**	**58**	**39.8**	**43.5**

Table 6.1: Voxel-wise classification results using: (a) Single Dictionary (1D), with 5000 atoms learned using healthy and the lesions class data, (b) 2-Class specific dictionaries with same size (2D-S): 5000 atoms each and (c) 2-Class specific dictionaries with different size (2D-D): 5000 atoms for the healthy class and 1000 atoms for the lesions class. Sensitivity, Positive Predictive Value (PPV) and Dice scores (%) are given for each method and each patient. The last row indicates the average for a particular method for all the patients.

	$T_O = 0.01$	$T_O = 0.1$	$T_O = 0.2$	$T_O = 0.3$	$T_O = 0.4$
PPV (%)	65.27	62.02	59.99	57.60	52.63
Sensitivity (%)	64.80	61.36	60.39	58.12	53.36

Table 6.2: Performance analysis for MS lesions detection using 2-class specific dictionaries with different size (2D-D) for each class, with 5000 atoms for healthy class dictionary and 1000 atoms for the lesions class dictionary.



(a) (b)

(c)

Figure 6.4: Classification results for Patient 8. For illustration purpose, one slice has been arbitrarily selected. True Positives are in red, False Positives are in cyan, False Negatives are in green. Methods (a), (b) and (c) are the same as in Table 6.1.

6.3.2.1 Extending the Training Dataset

We are aware that we do not have a very large population for training. Hence we investigated the incorporation of longitudinal database into our analysis by considering MR sequences at 3 time points (M_0, M_3 and M_6) for all the patients. As the lesions evolve over the course of time, it is fair to consider

that each new dataset will enrich our learning model. Thus, we modified the training data, for each patient, in two ways: (1) Data at time-points M_0 and M_3, with 26 datasets and (2) Data at time-points M_0, M_3 and M_3, with 39 datasets. However, the lesion detection experiments for the same test subjects, as in previous experiments, using class specific dictionaries with the sizes of 5000 and 1000 for healthy and the lesions class respectively, did not show any significant improvement in the sensitivity and PPV. This suggests that the population for training the dictionaries earlier was sufficient and the dictionaries should be adapted to learn more specific structures viz. WM, GM and CSF versus lesions to help improve the detection.

This experiment suggests that the classification approach using sparse representation and dictionary learning technique in such application is favorable with respect to other machine learning techniques that require much larger sets of training data. For example, the availability of huge amount of data was one of the main reasons behind the success of machine learning technique such as deep learning. However, the manual delineation of MS lesions is time consuming and requires experts. This limits the labeled training data that can be obtained in such application. From the above experiments, the dictionary learning proves to be effective in the compact representation of the data and achieves similar results even when large training data is not available.

6.3.3 Dictionary Size Selection

The selection of dictionary size for each class remains an important issue. As discussed in Section 5.2.1, we performed the following experiments in order to study how the dictionary size could be selected.

6.3.3.1 Principal Component Analysis of the Data

The PCA of the training data for the healthy brain tissues and the lesions class was used to find the number of eigenvectors required to reach the specified percentage of cumulative variance for each class. The results are shown in Table 6.3. Each entry along the first and the second row in the bale indicates the number of eigenvectors needed to reach 95%, 98% and 99% cumulative variance for the data corresponding to the healthy and the lesions classes, respectively.

Firstly, it can be observed that the number of eigenvectors required to attain a respective cumulative variance for the healthy class data is greater than the lesions class. This suggests that the data corresponding to the healthy brain tissues is associated with more variability as compared to the lesions class. This variability needs to be taken into account in the dictionary learn-

	95%	98%	99%
Healthy	63	143	208
Lesions	31	71	121

Table 6.3: Principal component analysis of the training data for an arbitrarily selected patient. For each class mentioned in a row, an entry in the table denotes the number of eigenvectors required to attain the percentage of total variance indicated in each column.

ing based classification. We consider this variability difference between class data by using different dictionary sizes for the healthy and the lesions classes. The requirement of larger number of eigenvectors for the healthy class data suggests the use of a larger dictionary size for the healthy class data. Our experiments confirm this fact from the comparison of the classification results for the same or different dictionary sizes for the healthy and the lesions class. The mean dice score using the same dictionary size of 5000 for both classes is 12.8%, which increases to 43.5% when the different dictionary sizes, 5000 for the healthy class and 1000 for the lesions class is used. This confirms the fact that PCA can be used to consider the variability differences between class data and subsequently use the different dictionary sizes for each class.

It must however be noted that the PCA did not give the exact ratio of dictionary sizes to be used for the optimal classification. The data for the healthy class required approximately twice the number of eigenvectors as compared to the lesions class. However, the optimal classification is achieved with the use of ratio of 5 for the dictionary size for the healthy and the lesions class. The inability of PCA to suggest the dictionary size might be because of the nonlinearities associated with the healthy class data. We will see in the next section, how this problem can be tackled when we will use the class data for individual healthy brain tissue, WM, GM and CSF, where each tissue follows the Gaussian distribution.

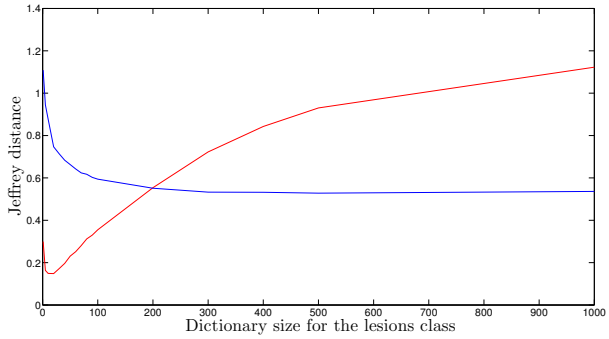
6.3.3.2 Histogram Based Measures

The variability differences between the data for the healthy and the lesions class is not taken into account if we incorporate the dictionaries of same size for both the classes. The simplest idea to consider the variability differences while performing classification is to learn the dictionaries for each class so that these dictionaries have the same level of representativity for both the classes. The average reconstruction error for the patches belonging to the healthy and the lesions class could be used for this purpose. By keeping the dictionary size for one of the classes fixed, the size of the dictionary for the other class

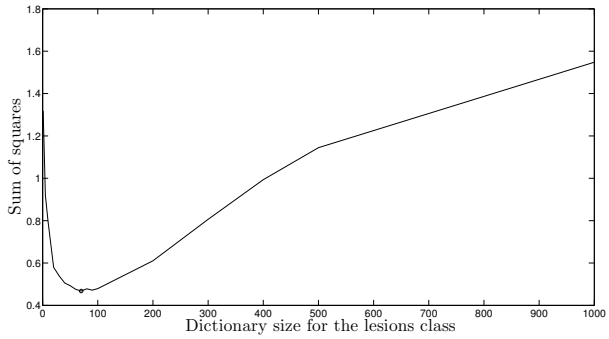
could be varied and the dictionary size for which the average reconstruction errors for both the classes match to each-other, can be selected as the optimal dictionary size. In our experiments, the average error did not prove to be sufficient measure and that is why we investigated more sophisticated measure based on histograms of the reconstruction errors, as described next.

The reconstruction errors for each class data obtained from the class specific dictionaries are analyzed to calculate the histogram based measures for selecting the dictionary sizes for each class, as discussed in Section 5.2.1.2. The objective is to fix the dictionary size for the healthy class and for various dictionary sizes for the lesions class, the histograms of reconstruction errors are obtained. The optimal dictionary size is found by matching the histograms corresponding to the reconstruction errors for the class data using the dictionary for the same classes and also the histograms belonging to the reconstruction errors for the class data using the opposite class dictionary. The first condition guarantees the same level of representativity for both classes using the dictionary for the same class, while the second condition is imposed for the opposite class dictionaries to be least representative for the given class data.

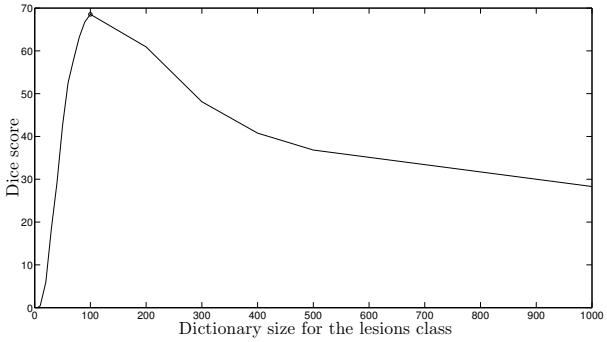
For the purpose of demonstration, we consider the simple case of using the same training and the test data for the patchwise classification of MS lesions for a randomly selected patient. As described in Section 5.2.1.2, Figure 6.5 (b) shows the sum of squares of the Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ and $d_{J2}(H_{1,2}, H_{2,1})$. It can be seen that the minimum value of the squared sum of these Jeffrey distance measures is obtained for the dictionary size of 1000 for the healthy class and 70 for the lesions class. The variation of dice score with respect to change in the dictionary size for the lesions class from 1 to 1000, while using the fixed dictionary size of 1000 for the healthy class data is shown in Figure 6.5 (c). The best dice score is obtained using the dictionary size of 100 for the lesions class. Although the dictionary size suggested by the histogram based measures does not exactly produce the best classification, the dice score using the dictionary size suggested by the proposed measure is still higher than the Dice score achieved using the same dictionary size 1000 for both classes. We observe the similar trend for all 13 MS patients for which this experiment was performed. In some cases, the histogram based method suggested the exact dictionary size for the lesions class for which optimal Dice score was obtained, while in other cases, the dictionary size suggested by this method did not deviate too far from the optimal dictionary size observed using the variation of Dice score for various dictionary sizes for the lesions class.



(a) Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ in red and $d_{J2}(H_{1,2}, H_{2,1})$ in blue, for the comparison of histograms



(b) Sum of squares of the Jeffrey divergence measures $d_{J1}(H_{1,1}, H_{2,2})$ and $d_{J2}(H_{1,2}, H_{2,1})$. The minimum value is achieved at the dictionary size of 70 for the lesions class, as indicated by the circled point on the curve.



(c) Dice scores for the blockwise classification of MS lesions. The best classification is obtained at the dictionary size of 100 for the lesions class, as indicated by the circled point on the curve.

Figure 6.5: The selection of dictionary size of the lesions class using histogram based measures. The dictionary size for the healthy class is kept constant as 1000 and the dictionary size for the lesions class is caried from 1 to 1000. The optimal dictionary size for the lesions class is chosen as 70, as indicated in Figure (b), where as the best classification result is obtained using the dictionary size of 100, as shown in Figure (c)

	100	200	1000	5000
SDL-S	7.0/13.0	8.4/15.3	12.8/22.1	13.4/23.2
FDDL-S	6.7/12.5	7.3/13.5	**14.0/24.0**	X

Table 6.4: Comparison of MS Lesion Classification using Standard Dictionary Learning (SDL) and Fisher Discrimination Dictionary Learning (FDDL) using the same dictionary size for the healthy and the lesions class: The results (PPV/Dice scores) of patch-wise classification for MS patient with High Lesion Load. 'X' indicates experiment not performed because of higher computation time requirement.

	100	200	1000	5000
SDL-D	14.5/20.8	25.9/31.6	40.4/44.3	61.0/49.6
FDDL-D	20.5/26.3	32.0/36.3	**61.7/55.1**	X

Table 6.5: Comparison of MS Lesion Classification using Standard Dictionary Learning (SDL) and Fisher Discrimination Dictionary Learning (FDDL) using different dictionary sizes for the healthy and the lesions class: The results (PPV/Dice scores) of patch-wise classification for MS patient with High Lesion Load. 'X' indicates experiment not performed because of higher computation time requirement.

6.3.4 Role of Dictionary Size in the Discriminative Dictionary Learning

As discussed in the previous chapters, the discriminative dictionary learning approaches have been proposed for improving the classification. We consider a particular discriminative dictionary learning technique called Fisher Discrimination Dictionary Learning (FDDL) and explore the role of the dictionary size in the case of MS lesions classification. The results of the classification using Standard Dictionary Learning (SDL) and FDDL are compared when we use the same or different dictionary sizes for the healthy brain tissue and the lesions class. The reader is referred to the Section 4.3 for the description of the method.

The classification experiments are first performed using SDL with the same and different dictionary sizes, respectively. These dictionaries are then used as an initialization in the dictionary learning step in the FDDL. The results obtained using FDDL are then compared with SDL method, for both the same and different dictionary sizes.

To experiment using FDDL, which is computationally inefficient when a large number of training samples are used, we sampled the training data for

	100	200	1000	5000
SDL-S	3.4/6.5	4.1/7.8	6.3/11.7	**9.7/17.3**
FDDL-S	3.1/6.0	3.5/6.8	X	X

Table 6.6: Comparison of MS Lesion Classification using Standard Dictionary Learning (SDL) and Fisher Discrimination Dictionary Learning (FDDL) using the same dictionary size for the healthy and the lesions class: The results (PPV/Dice scores) of patch-wise classification for MS patient with Low Lesion Load. 'X' indicates experiment not performed because of higher computation time requirement.

	100	200	1000	5000
SDL-D	11.6/18.4	21.5/23.3	32.2/36.6	**38.4/39.0**
FDDL-D	14.4/22.6	29.2/33.6	X	X

Table 6.7: Comparison of MS Lesion Classification using Standard Dictionary Learning (SDL) and Fisher Discrimination Dictionary Learning (FDDL) using different dictionary sizes for the healthy and the lesions class: The results (PPV/Dice scores) of patch-wise classification for MS patient with Low Lesion Load. 'X' indicates experiment not performed because of higher computation time requirement.

the healthy class by selecting 20K samples of healthy patches for each patient. We compared the results of classification with and without the sampling of training data, and found them to be very similar. However, FDDL method still required a large computation time. We, therefore, performed FDDL classification for two MS patients, one with a high lesion load and the other with a low lesion load. The results of classification using SDL and FDDL are shown in Tables 6.4, 6.5, 6.6, 6.7. It can be seen that using same dictionary size, the increase in dictionary sizes results in capturing more details and there is hence increase in both PPV and Dice values. However, the classification results improve drastically when we use the dictionaries of different sizes in the case of both SDL and FDDL. Moreover, FDDL with different dictionary size results in higher PPV and Dice scores than the corresponding SDL experiment with different dictionary size.

We performed these experiments on a high performance machine with 20 cores at 2.5 GHz and 128 GB of RAM. For the dictionary sizes from 100 to 1000, the classification using SDL took 5-12 minutes, whereas FDDL required 20-128 hours. Therefore, we did not perform FDDL classification experiments with higher dictionary size, marked as X in Tables 6.4, 6.5, 6.6, 6.7.

Figure 6.6 shows the best classification results obtained using methods discussed above. The classification image for FDDL with same dictionary size is similar to the one obtained using SDL with same dictionary size and is not shown in this figure. It can be seen that the method using same dictionary size results in many false positives, which are drastically reduced with the use of different dictionary sizes based on the variability of the class data.

6.4 MS Lesions Segmentation: 4-Class Method

There exist several MS lesions segmentation methods that use tissue segmentation to help segment the lesions [Zijdenbos 2002]. We can thus further enrich our model by taking into account the tissue specific information and learning dictionaries specific to different tissue types, such as White Matter (WM), Gray Matter (GM) and Cerebrospinal Fluid (CSF), as opposed to learning a single dictionary for healthy tissue patches. We explore the fact that various tissues as well as lesions appear in different intensity patterns in distinct MR modality images. For example, WM appears as the brightest tissue in T1-weighted image, but the darkest in T2-weighted images. Therefore, learning class specific dictionaries for individual tissues should further discriminate between lesion and non-lesion classes.

The dictionaries learned for each class are aimed at better representation of an individual class. However, if there exists differences in the data-complexity

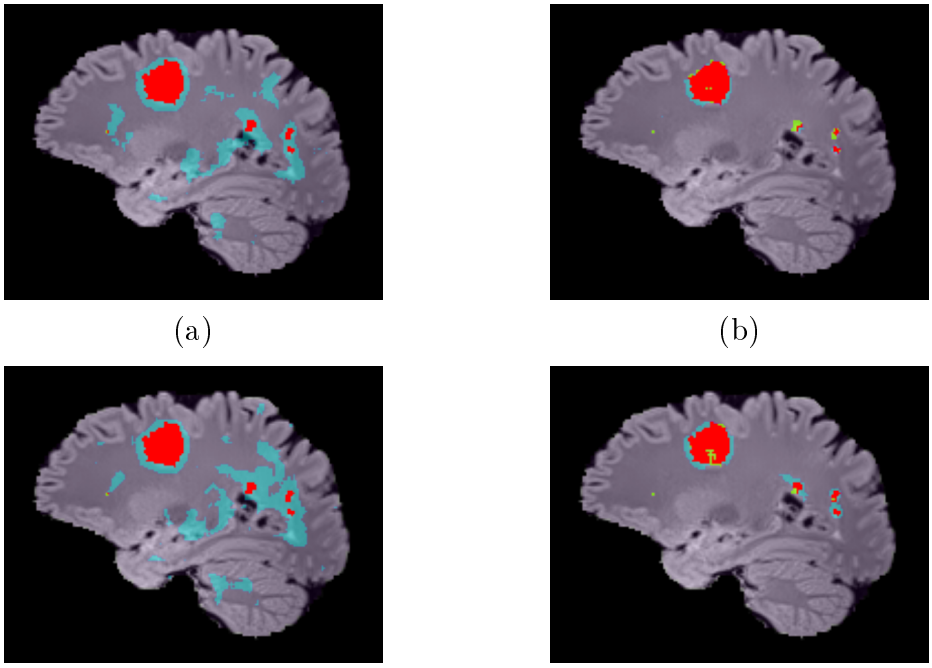


Figure 6.6: Classification results for MS patient with high lesion load, using (a) SDL with same dictionary size of 5000, (b) SDL with different dictionary sizes: 5000 for the healthy and 1000 for the lesion class, (c) FDDL with same dictionary size of 1000, and (d) FDDL with different dictionary sizes: 1000 for the healthy and 400 for the lesion class. The classification image is superimposed on FLAIR MRI. True Positives are in red, False Positives are in cyan, False Negatives are in green.

between classes, the relative under- or over-representation of either class will lead to worse classification. One idea for better classification could be to learn the dictionaries with adaptive sizes, in order to take into account the data variability between different classes. Thus, in addition to the dictionary learning strategy mentioned above, we also investigate the effect of modifying the dictionary sizes, leading to the proposition of adaptive dictionary learning. The basic idea is to learn the class specific dictionaries which are better adapted to the data and also complexity of the data.

The use of class specific dictionaries for each healthy brain tissue is also motivated from one of the observations in the previous method in selecting the dictionary size using PCA. As shown by the PCA analysis in Section 6.3.3.1, the class data for the healthy brain tissues is a rather non-linear data that can be explained by the different tissues embedded in this "meta" class. One

solution then is to learn the dictionaries specific to each healthy brain tissue in order to represent the sub-classes of the "healthy" tissues. It is well known that each healthy brain tissue follows the Gaussian distribution. In this manner, the disadvantage associated with the inability of PCA to handle non-linearity in the healthy class data can be avoided and the problem of dictionary size selection can be studied more efficiently.

The main contributions of this work can be outlined as follows: (1) Tissue-specific information is incorporated by learning dictionaries specific to each tissue class as opposed to learning a single dictionary for representation of the healthy brain tissue class, and (2) The dictionary sizes are adapted according to the complexity of the underlying data so that the dictionaries are better suited for representation of each class data as well as classification of MS lesions.

6.4.1 Overview of the method

The overview of the method proposed is shown in Figure 6.7. MR images for all patients are first preprocessed for noise-reduction and the elimination of extracranial brain tissues. The images are then registered into the same space. We represent image volumes as patches of a predefined size and normalize these extracted patches. This is followed by labeling patches in two ways: (i) Healthy brain tissue and the lesions patches, using manual segmentation images and (ii) WM, GM, CSF and the lesions patches, with the help of manual lesion segmentation and tissue segmentation images. The patches are then divided into the training and test dataset. For various classification strategies, we learn the dictionaries, using training data, in different configurations as follows: a single dictionary, two separate dictionaries for the healthy and lesion classes, or the class specific dictionaries for the lesions and each healthy brain tissue - WM, GM, CSF. For the last two approaches, we also study the role of the dictionary size in the classification. Finally, for a given test subject, we developed a reconstruction error based patch-classification method, which is followed by the voxel-wise classification. The following subsections briefly describe these steps.

6.4.1.1 Patch extraction and labeling

For local image analysis in the dictionary learning framework, the images are divided into the overlapping patches. Each patch is then represented as a signal in the dictionary learning process. We follow this patch-based approach and divide the whole intracranial MR volume for each patient into 3-D patches, with a patch around every 2 voxels in each direction. The individual image

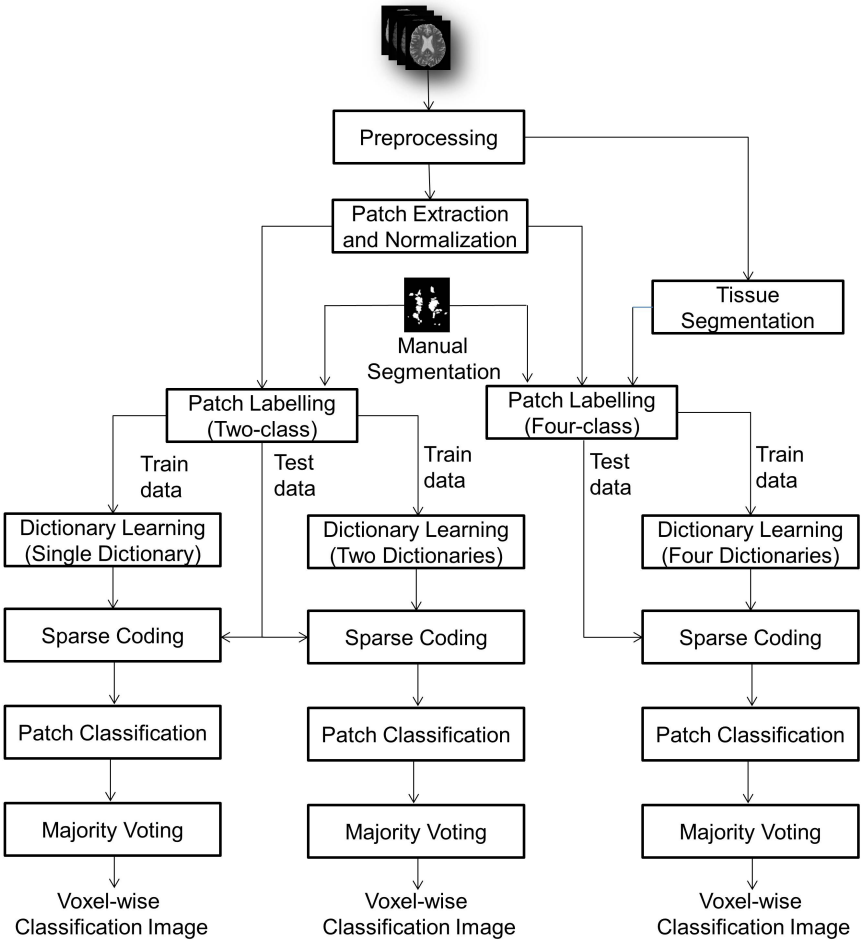


Figure 6.7: Flowchart of MS Lesions Classification using Dictionary Learning (4-Class Method)

patches of each MR modality are then flattened to form a vector and are concatenated together. The patches so obtained are normalized for a unit l_1 norm.

Next step is to label the normalized patches obtained from every patient. We label them in two different ways for the experiments to be preformed next. Firstly, the patches are labeled as belonging to either healthy or lesions class, using the manual segmentation image. If the number of lesion voxels in the corresponding image block of the manual segmentation image exceeds a pre-defined threshold T_L, we assign this patch to the lesions class. Otherwise, it is labeled as a healthy patch. The image patches obtained in this manner form the dataset for the classification approaches which use a single dictionary or two class specific dictionaries. For other classification methods, the patches are labeled as either WM, GM, CSF or the lesions class. We

use the same rule, as explained above, to label the patch to the lesions class. In addition, the patch is now assigned to either WM, GM or CSF class, depending on the maximum number of voxels that belong to corresponding class in the brain tissue segmentation image obtained using Statistical Parametric Mapping (SPM) [Ashburner 2005].

The labeled image patches are then divided into training and test data, and the experiments are performed by following Leave-One-Subject-Out-Cross-Validation (LOSOCV).

6.4.1.2 Patch-based classification using dictionary learning

Let n be the number of voxels per patch. For each class c, we write patches as vectors $\mathbf{x}_i^c \in R^n$. Learning an over-complete dictionary $D^c \in R^{n \times k}$ that is adapted to m patches, with sparsity parameter λ, is addressed by solving the optimization problem, similar to Equation 2.

$$\min_{D^c, \{\alpha_i^c\}_{i=1,..,m}} \sum_{i=1}^{m} \|\mathbf{x}_i^c - D^c \alpha_i^c\|_2^2 + \lambda \|\alpha_i^c\|_1 \tag{6.4}$$

The subsections below detail the different strategies adopted while learning these dictionaries and the scheme of patch based classification. In every method, we obtain the sparse codes for the test patches using Equation 6.1, knowing the dictionary D^c for the class c.

(a) Single Dictionary (1D)

As described in Section 6.3.1.3 (a), a single dictionary is learned from the healthy and the lesions class patches, and the lesions are classified as outliers when the reconstruction error for the test patch exceeds the chosen threshold [Weiss 2013].

(b) Two-Dictionaries: Same dictionary size (2D-S)

As described in Section 6.3.1.3 (b), the class specific dictionaries D^c of the same size are learned for the healthy ($c = 1$) and lesions ($c = 2$) classes. The classification is performed for a given test patch $\mathbf{y_i}$, by calculating the sparse coefficients α_i^c for each class and the test patch is then assigned to the class with a minimum representation error.

$$c_{pred} = \underset{c}{\operatorname{argmin}} \|\mathbf{y_i} - D^c \alpha_i^c\|_2^2 . \tag{6.5}$$

(c) Two-Dictionaries: Different dictionary size (2D-D)

The dictionaries of different sizes are learned for the healthy and the lesions class, in order to take into account the variability differences between class data, as described in Section 6.3.1.3 (c).

(d) Four-Dictionaries: Same dictionary size (4D-S)

As explained before, the healthy brain tissues contain anatomically different regions such as WM, GM and CSF. The fact that every tissue, WM, GM and CSF, appears in different intensity pattern in each MR modality, using a single dictionary for representing the healthy brain tissues might not be as effective as learning separate dictionaries for each tissue. Adding tissue specific information in the dictionaries used for the classification would enhance the prior knowledge in the learning step, thus highlighting the differences between individual tissues and also improving the lesion classification.

After learning class specific dictionaries for WM, GM, CSF and the lesions, we perform classification based on reconstruction error in a similar manner, as mentioned in method (b). Each dictionary is representative of its own class and the reconstruction of the test data using true class dictionary would give a minimum reconstruction error.

(e) Four-Dictionaries: Different dictionary size (4D-D)

Here, we experiment with different dictionary sizes for WM, GM, CSF and the lesions classes, for the similar reasons mentioned in method (c).

6.4.1.3 Voxel-wise classification

As already stated, we classify the patches centered around every 2 voxels in each direction. For voxel-wise classification, we assign each voxel to either of the classes by using majority voting. The voxel is assigned to a class using majority votes of all patches that contain the voxel.

Finally, in the context of lesion classification, we record the number of voxels that belong to True Positives (TP), False Negatives (FN) or False Positives (FP), and calculate percentage sensitivity (SEN)= $\frac{TP\times100}{TP+FN}$, percentage Positive Predictive Value (PPV) = $\frac{TP\times100}{TP+FP}$ and percentage dice-score (Dice) = $\frac{2\times TP\times100}{2\times TP+FP+FN}$.

6.4.2 Experiments and Results

For labeling patches, we used the threshold $T_L = 6$, as mentioned in Section 6.4.1.2. For patch size of 5×5×5, the number of lesion patches for each patient varied from 1K to 30K, depending on the lesion load for the corresponding patient, whereas the average number of patches for the healthy brain tissue class was 1.5×10^6. The brain tissue segmentation was obtained using Statistical Parametric Mapping (SPM) [Ashburner 2005]. The numbers of patches obtained per patient for WM, GM and CSF classes were 50K, 90K

and 30K, respectively. The classification was performed using LOSOCV and different parameters were tested. It was found that the patch size of 5×5×5 and the sparsity parameter $\lambda = 0.95$ were optimal choices. Changing λ in steps of 0.5 from 0.1 to 0.95 did not influence the results much and the value of 0.95 provided good results for all patients. All these experiments were performed on 2.5 GHz, 120 GB RAM Xeon processor. The dictionaries of sizes ranging from 500 to 5000, were learned from the training data and the best results, in terms of both sensitivity and PPV, were selected. For the dictionary sizes varying from 500 to 5000, the dictionary learning step required 5 minutes to 3 hours, where as the classification step took 4 minutes to 38 minutes, respectively. We used these parameters for validation of classification approaches using multi-channel MR data. We, however, excluded one patient with strong MR artifacts from this analysis.

The results of voxel-wise classification, obtained using all the methods described above, are shown in Table 6.8. Method (a) indicates classification obtained using single dictionary learned with the help of both healthy brain tissue and the lesions patches. Here, we chose the sparse penalty factor $\lambda = 0.85$ in the sparse coding step and performed the classification for various threshold values on the histogram of error map, as explained in Section 6.3.1.3 (a). The threshold, which produced the best voxel-wise classification results in terms of both sensitivity and PPV, was then selected and the classification results were reported. It can be observed from very low PPV and dice-scores that this method suffers with a very large number of false positive detections.

Pat. No.	(a) 1D			(b) 2D-S			(c) 2D-D			(d) 4D-S			(e) 4D-D		
	SEN	PPV	Dice	SEN	PPV	Dice	SEN	PPV	Dice	SEN	PPV	Dice	SEN	PPV	Dice
1	42	1	0.2	97	3	4.3	53	31	38.5	67	15	23.1	39	39	38.6
2	74	1	0.3	98	2	3.7	66	41	50.4	80	15	24.7	65	44	51.9
3	73	1	0.4	91	2	3	63	27	36.8	71	14	22.3	59	31	40.1
4	91	2	2.3	98	17	27.9	57	68	61.4	88	62	72.6	71	83	76.2
5	61	1	1.2	95	10	18	54	65	58.8	84	52	64	69	71	69.6
6	91	7	12.4	89	29	42.9	38	55	44.4	79	51	61.1	59	64	60.7
7	78	1	0.5	85	3	5.3	20	32	24.2	63	23	33.3	37	36	35.8
8	72	1	0.8	98	3	4.4	69	21	31.6	89	12	20.6	73	24	35.9
9	66	1	1.2	97	9	15.2	61	52	55.7	85	41	54.6	71	63	65.9
10	89	2	3.6	98	12	21.2	66	41	50.3	90	32	47	75	47	57
11	75	1	1.4	99	8	13.5	52	36	42.3	82	25	38	62	41	48.5
12	78	1	0.9	100	3	5.3	77	31	43.8	91	15	24.8	73	30	41.5
13	59	1	0.3	100	2	2.3	78	17	27	88	7	11.4	68	16	25.2
Mean	**73**	**1.6**	**2**	**95.8**	**7.9**	**12.8**	**58**	**39.8**	**43.5**	**81.3**	**28**	**38.3**	**63.2**	**45.3**	**49.8**

Table 6.8: Voxel-wise classification results using: (a) Single Dictionary, with 5000 atoms learned using the healthy and lesion class data, (b) Two class specific dictionaries with 5000 atoms each for the healthy and the lesion class, (c) 5000 atoms for the healthy and 1000 atoms for the lesion class dictionary, (d) Four class specific dictionaries with 5000 atoms each for WM, GM, CSF and the lesion classes, (e) 4000 atoms each for WM, GM and CSF classes, and 2000 atoms for the lesion class dictionary.

In the second experiment, we used the class specific dictionaries of same size, for the healthy and the lesions class. As indicated by method (b), the classification obtained using dictionaries with 5000 atoms each resulted in high sensitivity but PPV and dice-scores were still low. One possible reason behind these low values is that there exists a difference in variability of the data for two classes. Considering more variability associated with the healthy class data, we then used different dictionary sizes, 5000 for the healthy class and 1000 for the lesions class. As shown in method (c), this drastically reduced FP, improving PPV and dice-scores, but also decreased the sensitivity.

We further enriched this model by learning separate dictionaries for each healthy brain tissue - WM, GM, CSF, in addition to the dictionary learned for the lesions class. Using four such dictionaries with 5000 atoms each, it can be observed that a better compromise between sensitivity and PPV is achieved, as compared to methods (b) and (c) described above. This is shown by method (d). Finally, the classification using four dictionaries of different sizes, 4000 each for WM, GM and CSF classes, and 2000 for the lesions class, was obtained. This reduced the mean sensitivity but improved both the mean PPV and the mean dice-score, as compared to method (d) and is indicated by method (e) in Table 6.8.

The methods (c) and (e), which consider the inter-class data variability and use different dictionary sizes in classification, offer a better compromise between sensitivity and PPV, as compared to their counterpart methods (b) and (d), which use the same dictionary size for all classes. Between methods (c) and (e), each employing either two or four dictionaries respectively, the later method performs better than the former with a higher mean sensitivity, PPV and dice-score. Their comparison also shows a significant difference in PPV and dice-scores, with respective p-values of 0.0008 and 0.003. This confirms that the classification improves using dictionaries for each brain tissue.

6.4.2.1 Role of Dictionary Size on Classification

To investigate the effect of dictionary size on the performance of classification, we performed the experiments using methods (d) and (e) that use three separate dictionaries for the healthy brain tissues and one for the lesions class. Table 6.9 summarizes the results of classification.

For method (d), which uses the same dictionary size for all classes, the results along the diagonal of the table from top-left to bottom-right show that the sensitivity and PPV increase when the dictionary size is increased from 500 to 5000. The possible reason for this is that the dictionaries capture more details with the increase in their size. However, it can be observed that PPV values are very low for these experiments, indicating that this method suffers

with many false positive detections. Also, the increment is sensitivity and PPV values in very small when we increase the dictionary size from 500 to 5000 for all the classes simultaneously.

Excluding values along the diagonal mentioned above, all other entries in the table indicate the sensitivity and PPV values obtained with method (e), which uses different dictionary sizes for tissues and the lesions class. By referring to values in the columns from a single row, which suggests using a constant dictionary size for each tissue while varying the dictionary size of lesions class from 500 to 5000, we can observe that sensitivity keeps increasing but PPV value reduces, resulting in false positive detections. On the other hand, if we fix the dictionary size for the lesions class and increase the dictionary size for the tissues, PPV increases but sensitivity reduces, resulting in under-detection. Very low PPV scores above-diagonal from top-left to bottom-right suggest that the lesion dictionary over-represents the data corresponding to the lesion class, with the use of higher dictionary size for the lesions class than that for the tissue classes. The best results, for both sensitivity and PPV together, are obtained for the dictionary size of 4000 for each tissue class and 2000 for the lesions class. It can also be observed that it is the relative dictionary size that drives the classification and is more important than just the absolute dictionary size for each class.

	500	1000	2000	3000	4000	5000
500	*81.1 / 17.2*	94.9 / 5.2	98.6 / 2.5	99.2 / 2.2	99.4 / 2.2	99.5 / 2.1
1000	58.3 / 43.7	*81.8 / 19.7*	94.5 / 6.8	97.2 / 3.9	98.0 / 3.0	98.5 / 2.6
2000	32.7 / 65.2	60.9 / 44.1	*82.1 / 22.9*	89.8 / 13.3	93.4 / 8.8	95.3 / 6.5
3000	19.2 / 72.2	46.8 / 56.2	71.4 / 36.4	*82.1 / 25.1*	87.0 / 18.1	90.4 / 13.6
4000	12.7 / 76.2	36.9 / 63.3	63.2 / 45.3	75.2 / 34.1	*81.3 / 26.7*	85.5 / 21.3
5000	8.9 / 79.5	30.0 / 67.5	56.9 / 51.1	69.2 / 40.5	76.2 / 33.4	*81.3 / 28.0*

Table 6.9: Effect of dictionary size in voxel-wise classification of MS lesions. Each entry in the table indicates the sensitivity and PPV value for the MS lesions classification. The leftmost column indicates the dictionary size for each healthy tissue - WM, GM and CSF, whereas the topmost row indicates the dictionary size for the lesions class. The sensitivity and PPV values for each combination of dictionary size for the tissue and lesions classes are indicated in the corresponding entries of the table. The entries in italics on the diagonal of the table from top-left to bottom-right refer to method (d), which uses the same dictionary size for all classes, where as all other entries represent method (e) with different dictionary sizes for the healthy tissues and the lesions class.

It is crucial to adapt the size of the dictionaries to better control the classification. For such purpose, we analyzed the data using Principal Component

	95%	98%	99%
WM	46	106	167
GM	86	156	207
CSF	60	140	209
Healthy	63	143	208
Lesions	31	71	121

Table 6.10: Principal component analysis of the training data for an arbitrarily selected patient. For each class mentioned in a row, an entry in the table denotes the number of eigen-vectors required to attain the percentage of total variance indicated in each column.

Analysis (PCA), which gives an estimate of the intrinsic dimensionality of the data. Figure 6.8 shows the cumulative variance explained by the eigenvectors of different classes such as WM, GM, CSF, lesions and healthy. The number of eigenvectors required for explaining the mentioned percentages of the total variances for each class are shown in Table 6.10. It can be seen that, for each brain tissue - GM, WM and CSF, approximately twice as many eigenvectors are required for an arbitrary proportion of the percentage cumulative data variance (90%, 95% or 98%), as that required for the lesions data. As exhibited by method (e), this observation supports our adaption of dictionary size for each brain tissue twice that for the lesion dictionary. In case of method (c), which uses dictionaries for healthy and lesions classes, the experimentally observed optimal dictionary size ratio of 5 for the healthy and the lesions class was not found with PCA. Although, the factor 2 indicated by PCA still favors using a higher dictionary size for the healthy class. One reasoning behind this failure might be the inability of PCA to analyze the non-linearity in the data. The intrinsic dimensionality estimation for this highly non-linear data could be further point of investigation.

In Figures 6.9 and 6.10, we show the voxel-wise classification results obtained using all methods discussed above. We arbitrarily selected a slice for the patients 4 and 6, as referred to in Table 6.8. It can be seen from Figure 6.9-F that method (a) suffers with a large number of FP. The over-detections are reduced in methods (b) and (d), which use dictionaries of the same size for each class. This is indicated in Figures 6.9-G and 6.9-I, respectively. Methods (c) and (e) further improve the classification, as shown in Figures 6.9-H and 6.9-J, by employing the dictionaries of adapted sizes. However, the 2-class method (c) has many FN. As shown by method (e), including tissue specific information in such adaptive dictionary learning based approach results in significant improvement in the lesion classification with reduction in both FP

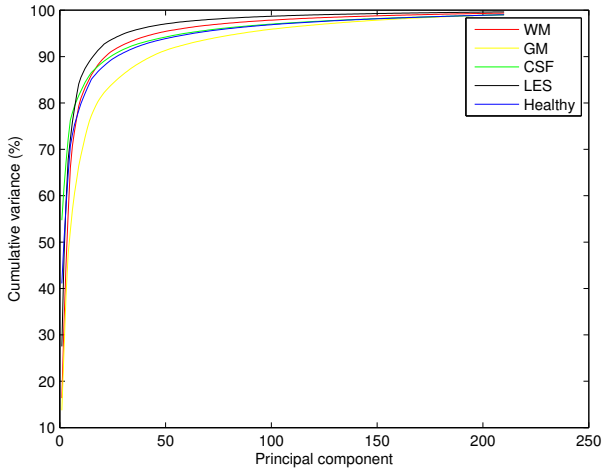


Figure 6.8: Cumulative variance for different classes, plotted against the number of principal components obtained from the principle component analysis of the corresponding class data.

and FN. This supports our claim that the method with the tissue specific dictionaries and adapted dictionary sizes is a better choice over the 2-class methods and those using the same dictionary size for all classes.

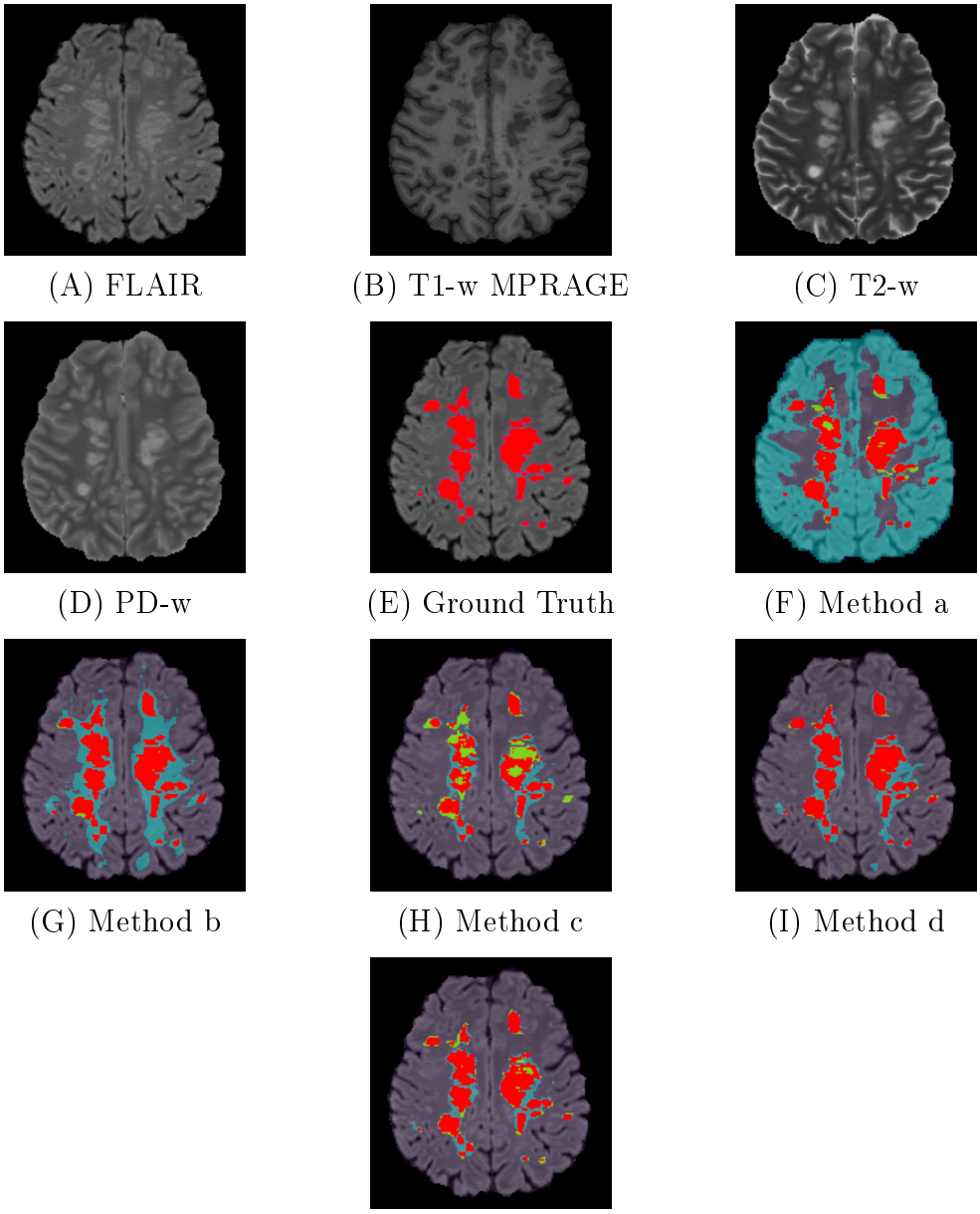
6.5 Conclusion

we proposed a new supervised approach to automatically detect multiple sclerosis lesions using dictionary learning. We investigated the performance of three methods which either use one dictionary, treating lesions as outliers, or use class specific dictionaries for healthy and the lesions classes, wherein the underlying data for each class is represented by the dictionary and sparse coefficients. We further studied the effect of using different dictionary sizes, allowing larger dictionaries to represent the complex data and concluded that such method minimizes the false positive detections in the classification. Although the method using class specific dictionaries follows supervised approach, contrary to the single dictionary based classification method, which does not necessarily require training data, it is worth mentioning that the former method eliminates one parameter: threshold on error map. This crucial paramater is not easy to tune and could lead to worse classification results for small errors in the brain extraction procedure.

Learning more specific dictionaries for each anatomical structure in the brain helps improve the classification on account of specific intensity patterns associated with each of these structures in multi-channel MR images. We

also demonstrated the effectiveness of adapting the dictionary sizes for better amplification of differences among multiple classes, hence improving the classification. If performing PCA on input data can successfully adapt the dictionary size for the classification, it is not as much efficient when the classes represent more a mixture of different tissues. Knowing the limitation of PCA to handle only linear data, future work could be to use the intrinsic dimension estimation techniques, which can better analyze complexity of the non-linear data.

We also evaluated the performance of the discriminative DL technique in the classification of MS lesions where the training data is complex and large in size, as compared to the computer vision applications such as face recognition or texture classification, which are used for validation by the sparsity community. The dictionary size played a major role even in the discriminative DL method such as Fisher Discrimination Dictionary Learning (FDDL). It was also found out that FDDL mechanism exhibits time-complexity issues in dealing with large data sets, as in medical imaging applications. Therefore, while dealing with pattern recognition in medical imaging, we strongly recommend to prefer DL methods that 1) can cope with the large size of medical images, and 2) that can adapt the size of the dictionaries according to the respective complexity of the patterns to detect.



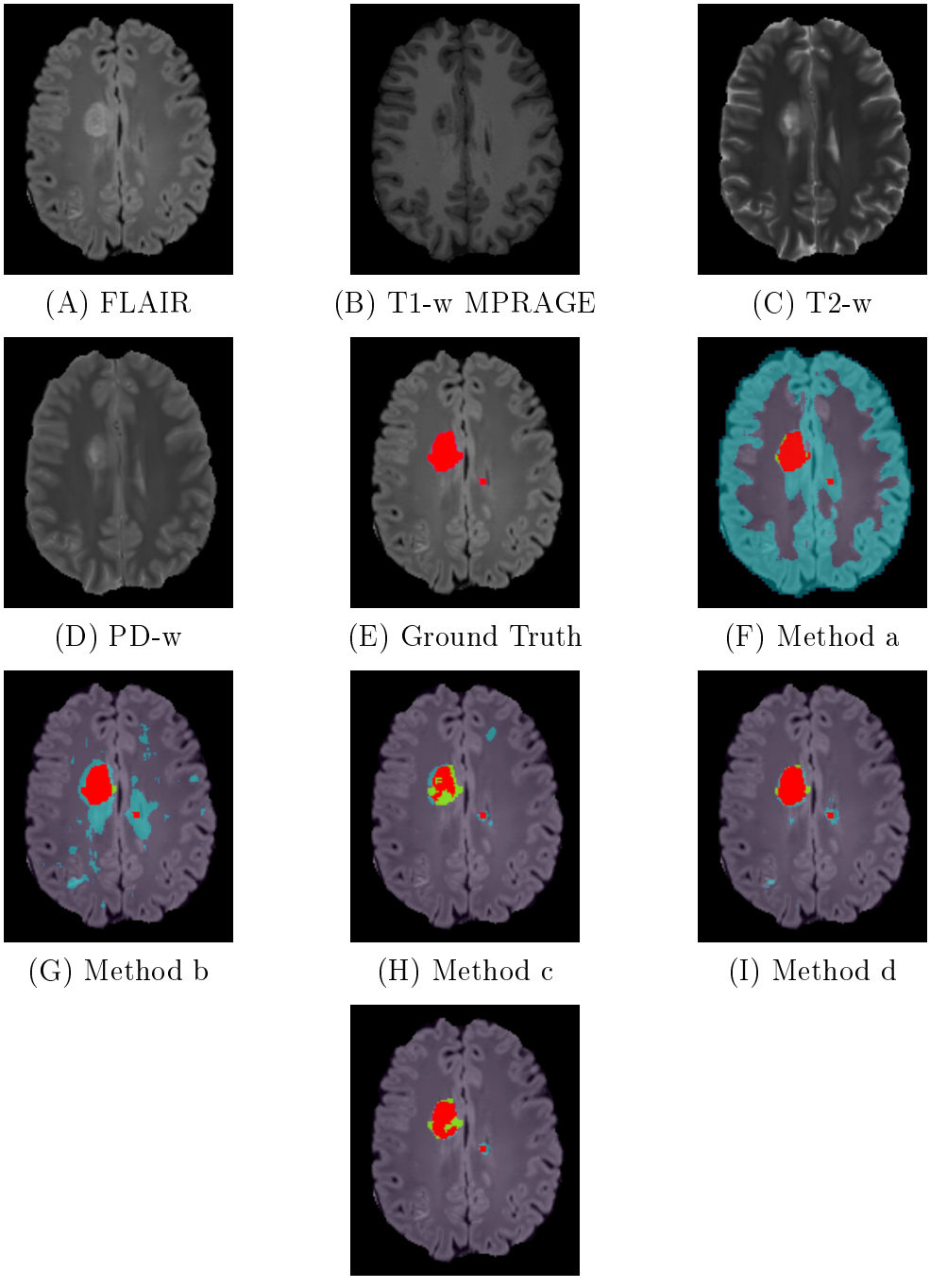
(A) FLAIR (B) T1-w MPRAGE (C) T2-w

(D) PD-w (E) Ground Truth (F) Method a

(G) Method b (H) Method c (I) Method d

(J) Method e

Figure 6.9: Comparison of MS lesion classification methods, example 1 - patient 6, slice 164. (A) FLAIR, (B) T1-w MPRAGE, (C) T2-w, (D) PD-w, (E) Ground truth or manual lesion segmentation image (shown in red) superimposed on FLAIR, (F) Result for method (a) using single dictionary with 5000 atoms learned using the healthy and lesions class data, (G) Result for method (b) with two dictionaries containing 5000 atoms each for the healthy and the lesions class, (H) Result for method (c) with 5000 atoms for the healthy and 1000 atoms for the lesions class dictionary, (I) Result for method (d) with four class specific dictionaries with 5000 atoms each for WM, GM, CSF and the lesions classes, (J) Result for method (e) with 4000 atoms each for WM, GM and CSF class, and 2000 atoms for the lesions class dictionary. Classification image is overlayed on FLAIR MRI. Red: TP; Cyan: FP; Green: FN.



(A) FLAIR (B) T1-w MPRAGE (C) T2-w

(D) PD-w (E) Ground Truth (F) Method a

(G) Method b (H) Method c (I) Method d

(J) Method e

Figure 6.10: Comparison of MS lesion classification methods, example 2 - patient 4, slice 153. (A) FLAIR, (B) T1-w MPRAGE, (C) T2-w, (D) PD-w, (E) Ground truth or manual lesion segmentation image (shown in red) superimposed on FLAIR, (F)-(J) Results of voxel-wise classification obtained using methods (a)-(e), as mentioned in Figure 6.9. Classification image is overlayed on FLAIR MRI. Red: TP; Cyan: FP; Green: FN.

Chapter 7

Conclusion

Contents

In this thesis, we have presented a pattern recognition framework using sparse representations and dictionary learning (DL) paradigm. The use of inherent sparsity property in most natural signals and learning relevant basis functions or a dictionary from the underlying data has led to interesting image representation and classification results, and remains an active research problem in the signal processing community. Incorporating these methods in medical imaging applications has additional challenges such as dealing with high complexity data and developing computationally efficient algorithms.

In conclusion, we now summarize the contributions made in this thesis and then discuss the perspective for future work.

7.1 Contributions

The dictionary learning has been used in several image processing applications such as denoising, inpainting, restoration, classification etc. We investigated the use of sparse representations and dictionary learning approach in pattern classification approaches where there are variability differences between patterns of interest and the background information.

First, we showed that the dictionary size for each class plays a major role in pattern classification with an example of computer vision application such as lips detection in face images. A prior information on variability differences between less complex lips data and more complex non-lips data is effectively used in the dictionary learning framework by incorporating different dictionary sizes for each class. We emphasize the fact that the dictionary size is not just a parameter in the dictionary learning framework, but it signifies two important properties of the dictionaries used in the classification: data

representation power and the inter-class discrimination ability. For the selection of dictionary size for optimal classification, we studied three different approaches: (i) PCA: The data complexity differences between class data are studied using the number of eigenvectors required to reach a particular value of cumulative variance for each class data. (ii) Histogram based measures: The dictionaries learned for each class are analyzed to obtain the histograms of reconstruction errors and the optimal dictionary size is selected when same level of representativity is attained for each class using the dictionaries of the same and the opposite class, and (iii) Empirical selection of dictionary size for each class for achieving the best classification.

Second, we proposed a supervised approach for the classification of Multiple Sclerosis (MS) lesions in multi-channel MR images. This is achieved by learning the class specific dictionaries for the healthy brain tissues and the lesions class, and allowing different dictionary sizes for each class for taking into account the variability differences between MS lesions and more complex healthy brain tissues. This method addressed two limitations of the previously proposed MS lesions segmentation approach using dictionary learning in unsupervised manner: (i) multi-channel MR images are employed in order to effectively utilize the contrast differences between healthy brain tissues and lesions, and (ii) a parameter which could led to worse segmentation for small errors in brain extraction process is eliminated to minimize the impact of preprocessing steps. We further discussed the problem of dictionary size selection using PCA and histogram based measures. We observed that PCA was unable to indicate the ratio of dictionary size for the two classes, supposedly because of the non-linear structures present in the healthy class data.

Third, the problem of dictionary size selection was addressed by reducing the non-linearity associated with healthy brain tissues. The dictionaries were learned for each healthy brain tissue - white matter, grey matter and cerebrospinal fluid, instead of learning a single dictionary for the combined class. This enriched the previous model, resulting in improved MS lesions segmentation performance and the underlying Gaussian distributions of each healthy brain tissue allowed the PCA to suggest the range of dictionary size for each class in order to achieve the best classification.

Fourth, the role of dictionary size in one of the most popular discriminative dictionary learning approaches - Fisher Discrimination Dictionary Learning (FDDL) - was investigated in the case of both: lips detection in face images and MS lesions classification. The addition of complex discriminative terms in the dictionary learning formulation was found to be less effective if the same dictionary size is used for each class. On the contrary, the different dictionary size for each class drastically improved the classification, suggesting the significance of dictionary size even in the case of discriminative dictionary

learning methods. One of the major disadvantages of this method was its high computational complexity, which further limited its use in the complex applications such as medical imaging.

The publications emerged from this work, until now, are as follows:

1. *Hrishikesh Deshpande*, Pierre Maurel, Christian Barillot, Classification of Multiple Sclerosis Lesions using Adaptive Dictionary Learning, Special Issue on Sparsity Techniques in Medical Imaging, Journal of Computerized Medical Imaging and Graphics, Elsevier, December 2015.
2. *Hrishikesh Deshpande*, Pierre Maurel, Christian Barillot, Adaptive Dictionary Learning For Competitive Classification Of Multiple Sclerosis Lesions, IEEE International Symposium on Biomedical Imaging (ISBI), New York, USA, April 2015.
3. *Hrishikesh Deshpande*, Pierre Maurel, Christian Barillot, Detection of Multiple Sclerosis Lesions using Sparse Representations and Dictionary Learning, 2nd Worskshop on Sparsity Techniques in Medical Imaging (STMI), 17th MICCAI, MIT, Boston, USA, September 2014.

7.2 Discussions and Future Work

The objective of this thesis was to investigate the use of sparse representation modelling, along with the dictionary learning techniques, in the classification of patterns in general and MS lesions in particular. While the first results provided on multi-sequence MR data are promising, it would be of great interest to take into consideration the lesion load information while developing a dedicated application for MS lesions classification using this technique. It was observed that parameters such as patch size and the dictionary size for the lesions class could be more effectively tuned for different values of lesion loads.

Our data set for MS lesions classification was confined to the use of T1-w MPRAGE, T2-w, PD-w and FLAIR sequences. It was observed that the combination of all these MR sequences lead to better performance when compared with the reduced data set consisting of few of these sequences. This suggests that the contrast information in each sequence adds discrimination information in the MS lesion classification using dictionary learning approach. Over the past years, the Gadolinium enhance T1-w MR imaging and quantitative MR sequences, such as DTI, MTR or even relaxometry, have also shown good sensitivity in the detection of MS lesions. It would be interesting to extend the proposed approach using these additional MR modalities.

MS lesions occur in different sizes, shapes and intensity patterns. In our approaches, we only considered the intensity values within patches of predefined size. Another possible direction is to extend this framework by experimenting with relevant features such as scale and rotational invariant features for the classification of MS lesions.

As discussed in previous section, the selection of dictionary size using PCA has inherent disadvantage if the underlying data is non-linear or non-Gaussian. One of the ways to tackle this problem would be to consider other approaches for quantifying the variability differences between the class data, for example, non-linear PCA, dimensionality estimation techniques etc.

The dictionary learning approaches have found applications in activity recognition, where a sequence of images is analyzed for detecting activities based on intensity differences between consecutive frames. In the case of MS, longitudinal studies are conducted to monitor disease progression and treatment efficiency. The MR images are analyzed for tracking the appearing or vanishing lesions. The dictionary learning approaches could be developed to detect such evolving lesions, instead of classifying just static lesions, as proposed in our approaches. Other possible future work could be to explore the role of sparsity techniques in the classification of other brain pathologies such as stroke or tumors.

Finally, several discriminative dictionary learning approaches have been proposed over the past few years, but they are mainly validated using computer vision applications. There are very few discriminative dictionary learning methods for medical imaging applications. One of the main disadvantages of these methods is computational complexity arising from high-dimensionality of the medical images. The development of discriminative dictionary learning methods which either scale to such high-dimensional data or extraction of low-dimensional relevant features which would speed up the performance of these methods would be another interesting future direction.

Bibliography

[Abdullah 2011] B. Abdullah, A. Younis, P. Pattany and E. Saraf-Lavi. *Textural Based SVM for MS Lesion Segmentation in FLAIR MRIs.* Open Journal of Medical Imaging, vol. 1, no. 2, pages 26–42, 2011. (Cited on page 80.)

[Aharon 2006] M. Aharon, M. Elad and A. Bruckstein. *K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation.* Trans. Sig. Proc., vol. 54, no. 11, pages 4311–4322, November 2006. (Cited on pages 30, 36 and 44.)

[Aït-Ali 2005] L. S. Aït-Ali, S. Prima, P. Hellier, B. Carsin, G. Edan and C. Barillot. Medical image computing and computer-assisted intervention – miccai 2005: 8th international conference, palm springs, ca, usa, october 26-29, 2005, proceedings, part i, chapter STREM: A Robust Multidimensional Parametric Method to Segment MS Lesions in MRI, pages 409–416. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. (Cited on page 83.)

[Akselrod-Ballin 2009] A. Akselrod-Ballin, M. Galun, J. M. Gomori, M. Filippi, P. Valsasina, R. Basri and A. Brandt. *Automatic Segmentation and Classification of Multiple Sclerosis in Multichannel MRI.* IEEE Transactions on Biomedical Engineering, vol. 56, no. 10, pages 2461–2469, Oct 2009. (Cited on page 80.)

[Anbeek 2004] Petronella Anbeek, Koen L. Vincken, Matthias J.P. van Osch, Robertus H.C. Bisschops and Jeroen van der Grond. *Probabilistic segmentation of white matter lesions in {MR} imaging.* NeuroImage, vol. 21, no. 3, pages 1037 – 1044, 2004. (Cited on page 81.)

[Anbeek 2005] Petronella Anbeek, Koen L. Vincken, Glenda S. van Bochove, Matthias J.P. van Osch and Jeroen van der Grond. *Probabilistic segmentation of brain tissue in {MR} imaging.* NeuroImage, vol. 27, no. 4, pages 795 – 804, 2005. (Cited on page 81.)

[Ashburner 2005] J. Ashburner and K. Friston. *Unified segmentation.* NeuroImage, vol. 26, no. 3, pages 839 – 851, 2005. (Cited on pages 102 and 103.)

[Baraniuk 2010] R. G. Baraniuk, E. Candes, M. Elad and Yi Ma. *Applications of Sparse Representation and Compressive Sensing [Scanning the*

Issue]. Proceedings of the IEEE, vol. 98, no. 6, pages 906–909, June 2010. (Cited on page 28.)

[Barkhof 1997] F Barkhof, M Filippi, D H Miller, P Scheltens, A Campi, C H Polman, G Comi, H J Adèr, N Losseff and J Valk. *Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis.* Brain, vol. 120, no. 11, pages 2059–2069, 1997. (Cited on page 79.)

[Brosch 2016] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee and R. Tam. *Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation.* IEEE Transactions on Medical Imaging, vol. 35, no. 5, pages 1229–1239, May 2016. (Cited on page 82.)

[Brunelli 2009] Roberto Brunelli. Template matching techniques in computer vision: Theory and practice. Wiley Publishing, 2009. (Cited on page 25.)

[Bryt 2008] Ori Bryt and Michael Elad. *Compression of facial images using the K-SVD algorithm.* Journal of Visual Communication and Image Representation, vol. 19, no. 4, pages 270 – 282, 2008. (Cited on page 41.)

[Burges 1998] Christopher J.C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, vol. 2, no. 2, pages 121–167, 1998. (Cited on page 17.)

[Cabezas 2013] Mariano Cabezas, Arnau Oliver, Jordi Freixenet and Xavier Lladó. Pattern recognition and image analysis: 6th iberian conference, ibpria 2013, funchal, madeira, portugal, june 5-7, 2013. proceedings, chapter A Supervised Approach for Multiple Sclerosis Lesion Segmentation Using Context Features and an Outlier Map, pages 782–789. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. (Cited on page 82.)

[Chen 1998] Scott Shaobing Chen, David L. Donoho, Michael and A. Saunders. *Atomic decomposition by basis pursuit.* SIAM Journal on Scientific Computing, vol. 20, pages 33–61, 1998. (Cited on pages 28, 31 and 86.)

[Chung 2011] M.K. Chung, Hyekyoung Lee, P.T. Kim and Jong Chul Ye. *Sparse topological data recovery in medical images.* In IEEE Interna-

tional Symposium on Biomedical Imaging: From Nano to Macro, pages 1125–1129, March 2011. (Cited on page 84.)

[Coupe 2008] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann and C. Barillot. *An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images.* IEEE Transactions on Medical Imaging, vol. 27, no. 4, pages 425–441, April 2008. (Cited on pages 80, 83 and 88.)

[Courtney 2006] Susan Wells Courtney. All about multiple sclerosis: Third edition. 2006. (Cited on page 75.)

[Deka 2010] B. Deka and P.K. Bora. *Despeckling of medical ultrasound images using sparse representation.* In International Conference on Signal Processing and Communications (SPCOM), pages 1–5, July 2010. (Cited on page 84.)

[Devroye 1996] Luc Devroye, Laszlo Gyorfi and Gabor Lugosi. A probabilistic theory of pattern recognition. Applications of mathematics. Springer, New York, Berlin, Heidelberg, 1996. Autres tirages : 1997, 2009. (Cited on page 24.)

[Donoho 2001] D. L. Donoho and X. Huo. *Uncertainty principles and ideal atomic decomposition.* IEEE Transactions on Information Theory, vol. 47, no. 7, pages 2845–2862, Nov 2001. (Cited on page 33.)

[Efron 2004] Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani. *Least angle regression.* Annals of Statistics, vol. 32, pages 407–499, 2004. (Cited on page 32.)

[Elad 2006a] M. Elad. *Why Simple Shrinkage Is Still Relevant for Redundant Representations?* IEEE Transactions on Information Theory, vol. 52, no. 12, pages 5559–5569, Dec 2006. (Cited on page 32.)

[Elad 2006b] M. Elad and M. Aharon. *Image Denoising Via Learned Dictionaries and Sparse representation.* In CVPR, volume 1, pages 895–900, June 2006. (Cited on pages 41 and 43.)

[Elad 2006c] M. Elad and M. Aharon. *Image Denoising Via Learned Dictionaries and Sparse representation.* In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 895–900, June 2006. (Cited on page 84.)

[Elad 2010] M. Elad, M. A. T. Figueiredo and Y. Ma. *On the Role of Sparse and Redundant Representations in Image Processing.* Proceedings of the IEEE, vol. 98, no. 6, pages 972–982, June 2010. (Cited on pages 30 and 84.)

[Elhamifar 2012] E. Elhamifar, G. Sapiro and R. Vidal. *See all by looking at a few: Sparse modeling for finding representative objects.* In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 1600–1607, June 2012. (Cited on page 43.)

[Engan 1999a] K. Engan, S. O. Aase and J. Hakon Husoy. *Method of optimal directions for frame design.* In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, volume 5, pages 2443–2446, 1999. (Cited on page 35.)

[Engan 1999b] Kjersti Engan, Bhaskar D Rao and Kenneth Kreutz-Delgado. *Frame design using FOCUSS with method of optimal directions (MOD).* Proc. NORSIG, vol. 99, pages 65–69, 1999. (Cited on page 35.)

[Espinosa 2014] Irlanda J. Espinosa. *Espinosa: Cause of Multiple Sclerosis and Other Demyelinating Diseases Given New Explanation By Researchers.* http://wwwmsviewsandrelatednews.blogspot.fr/2014/03/cause-of-multiple-sclerosis-other.html, 2014. Accessed: 10th May 2016. (Cited on page 74.)

[Fang 2013a] Ruogu Fang, Tsuhan Chen and Pina C Sanelli. *Tissue-specific sparse deconvolution for low-dose CT perfusion.* International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 16, no. 1, pages 114 – 121, 2013. (Cited on page 84.)

[Fang 2013b] Ruogu Fang, Tsuhan Chen and Pina C. Sanelli. *Towards robust deconvolution of low-dose perfusion CT: Sparse perfusion deconvolution using online dictionary learning.* Medical Image Analysis, vol. 17, no. 4, pages 417 – 428, 2013. (Cited on page 84.)

[Fang 2014] Ruogu Fang, Kolbeinn Karlsson, Tsuhan Chen and Pina C. Sanelli. *Improving low-dose blood-brain barrier permeability quantification using sparse high-dose induced prior for Patlak model.* Medical Image Analysis, vol. 18, no. 6, pages 866 – 880, 2014. Sparse Methods for Signal Reconstruction and Medical Image Analysis. (Cited on page 84.)

[Ferrari 2003] Ricardo J. Ferrari, Xingchang Wei, Yunyan Zhang, James N. Scott and J. R. Mitchell. *Segmentation of multiple sclerosis lesions using support vector machines.* volume 5032, pages 16–26, 2003. (Cited on page 81.)

[Fiot 2013] Jean-Baptiste Fiot, Laurent D. Cohen, Parnesh Raniga and Jurgen Fripp. *Efficient brain lesion segmentation using multi-modality tissue-based feature selection and support vector machines.* International Journal for Numerical Methods in Biomedical Engineering, vol. 29, no. 9, pages 905–915, 2013. (Cited on page 81.)

[Gao 2014] Shenghua Gao, Ivor Wai-Hung Tsang and Yi Ma. *Learning Category-Specific Dictionary and Shared Dictionary for Fine-Grained Image Categorization.* IEEE Trans. Image Processing, vol. 23, no. 2, pages 623–634, 2014. (Cited on pages 47 and 84.)

[García-Lorenzo 2008] Daniel García-Lorenzo, Sylvain Prima, D. Louis Collins, Douglas L. Arnold, Sean Patrick Morrissey and Christian Barillot. *Combining Robust Expectation Maximization and Mean Shift algorithms for Multiple Sclerosis Brain Segmentation.* In MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues) (MIAMS'2008), pages 82–91, New York, United States, September 2008. (Cited on page 83.)

[García-Lorenzo 2013] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L. Arnold and D. Louis Collins. *Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging.* Medical Image Analysis, vol. 17, no. 1, pages 1 – 18, 2013. (Cited on page 81.)

[Georghiades 2001] A. S. Georghiades, P. N. Belhumeur and D. J. Kriegman. *From few to many: illumination cone models for face recognition under variable lighting and pose.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pages 643–660, Jun 2001. (Cited on page 45.)

[Geremia 2010] Ezequiel Geremia, Bjoern H. Menze, Olivier Clatz, Ender Konukoglu, Antonio Criminisi and Nicholas Ayache. Medical image computing and computer-assisted intervention – miccai 2010: 13th international conference, beijing, china, september 20-24, 2010, proceedings, part i, chapter Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel MR Images, pages 111–118. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. (Cited on page 82.)

[Ghahramani 2004] Zoubin Ghahramani. Advanced lectures on machine learning: Ml summer schools 2003, canberra, australia, february 2 - 14, 2003, tübingen, germany, august 4 - 16, 2003, revised lectures, chapter Unsupervised Learning, pages 72–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. (Cited on page 18.)

[Goldberg-Zimring 1998] D Goldberg-Zimring, A Achiron, S Miron, M Faibel and H Azhari. *Automated Detection and Characterization of Multiple Sclerosis Lesions in Brain {MR} Images.* Magnetic Resonance Imaging, vol. 16, no. 3, pages 311 – 318, 1998. (Cited on pages 23 and 82.)

[Goldberg-Zimring 2005] Daniel Goldberg-Zimring, Andrea U. J. Mewes, Mahnaz Maddah and Simon K. Warfield. *Diffusion Tensor Magnetic Resonance Imaging in Multiple Sclerosis.* Journal of Neuroimaging, vol. 15, pages 68S–81S, 2005. (Cited on page 77.)

[Gorodnitsky 1997] I. F. Gorodnitsky and B. D. Rao. *Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm.* IEEE Transactions on Signal Processing, vol. 45, no. 3, pages 600–616, Mar 1997. (Cited on page 32.)

[Grosse 2012] Roger B. Grosse, Rajat Raina, Helen Kwong and Andrew Y. Ng. *Shift-Invariance Sparse Coding for Audio Classification.* CoRR, vol. abs/1206.5241, 2012. (Cited on page 41.)

[Grossman 1998] Robert Grossman and Joseph McGowan. *Perspectives on multiple sclerosis.* American journal of neuroradiology, vol. 19, no. 7, pages 1251–1265, Aug 1998. (Cited on page 76.)

[Harmouche 2006] R. Harmouche, L. Collins, D. Arnold, S. Francis and T. Arbel. *Bayesian MS Lesion Classification Modeling Regional and Local Spatial Information.* In 18th International Conference on Pattern Recognition (ICPR'06), volume 3, pages 984–987, 2006. (Cited on page 82.)

[Jain 1996] A. K. Jain, Jianchang Mao and K. M. Mohiuddin. *Artificial neural networks: a tutorial.* Computer, vol. 29, no. 3, pages 31–44, Mar 1996. (Cited on page 25.)

[Jiang 2011] Zhuolin Jiang, Zhe Lin and L. S. Davis. *Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD.* In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11, pages 1697–1704, Washington, DC, USA, 2011. IEEE Computer Society. (Cited on page 45.)

[Kamber 1995] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis and A. C. Evans. *Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images.* IEEE Transactions on Medical Imaging, vol. 14, no. 3, pages 442–453, Sep 1995. (Cited on page 82.)

[Kandel 2013] Benjamin M. Kandel, David A. Wolk, James C. Gee and Brian Avants. *Predicting Cognitive Data from Medical Images Using Sparse Linear Regression.* In Information Processing in Medical Imaging, volume 7917 of *Lecture Notes in Computer Science*, pages 86–97. Springer Berlin Heidelberg, 2013. (Cited on page 84.)

[Karpate 2014] Yogesh Karpate, Olivier Commowick, Christian Barillot and Gilles Edan. *Longitudinal Intensity Normalization in Multiple Sclerosis Patients.* In MICCAI Workshop on Clinical Image-based Procedures, pages 1–8, Boston, United States, September 2014. (Cited on page 80.)

[Karpate 2015] Yogesh Karpate, Olivier Commowick and Christian Barillot. *Probabilistic One Class Learning for Automatic Detection of Multiple Sclerosis Lesions.* In IEEE International Symposium on Biomedical Imaging (ISBI), pages 486–489, Brooklyn, United States, April 2015. (Cited on page 82.)

[Kasiński 2008] Andrzej Kasiński*et al. The PUT Face Database.* Image Processing & Communication, vol. 13, no. 3, pages 59 – 64, 2008. (Cited on page 56.)

[Khayati 2008] Rasoul Khayati, Mansur Vafadust, Farzad Towhidkhah and Massood Nabavi. *Fully automatic segmentation of multiple sclerosis lesions in brain {MR} {FLAIR} images using adaptive mixtures method and markov random field model.* Computers in Biology and Medicine, vol. 38, no. 3, pages 379 – 390, 2008. (Cited on pages 80 and 83.)

[Kotsiantis 2007] S. B. Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques.* In Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press. (Cited on page 17.)

[Kroon 2008] D.J. Kroon, E.S.B. van Oort and C.H. Slump. *Multiple Sclerosis Detection in Multispectral Magnetic Resonance Images with Principal Components Analysis.* In 3D Segmentation in the Clinic: A Grand

Challenge II: MS lesion segmentation, Website, September 2008. Kitware. (Cited on page 82.)

[Leemput 2001] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester and P. Suetens. *Automated segmentation of multiple sclerosis lesions by model outlier detection.* IEEE Transactions on Medical Imaging, vol. 20, no. 8, pages 677–688, Aug 2001. (Cited on page 83.)

[Lesage 2005] S. Lesage, R. Gribonval, F. Bimbot and L. Benaroya. *Learning unions of orthonormal bases with thresholded singular value decomposition.* In Acoustics, Speech, and Signal Processing, 2005. Proceedings. IEEE International Conference on, volume 5, pages 293–296, March 2005. (Cited on page 37.)

[Li 2012] Shutao Li, Leyuan Fang and Haitao Yin. *An Efficient Dictionary Learning Algorithm and Its Application to 3-D Medical Image Denoising.* Biomedical Engineering, IEEE Transactions on, vol. 59, no. 2, pages 417–427, Feb 2012. (Cited on page 41.)

[Lladó 2012] Xavier Lladó, Arnau Oliver, Mariano Cabezas, Jordi Freixenet, Joan C. Vilanova, Ana Quiles, Laia Valls, Lluis Ramio-Torrenta and Alex Rovira. *Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches.* Information Sciences, vol. 186, no. 1, pages 164 – 185, 2012. (Cited on page 81.)

[Maintz 1998] J.B.Antoine Maintz and Max A. Viergever. *A survey of medical image registration.* Medical Image Analysis, vol. 2, no. 1, pages 1 – 36, 1998. (Cited on page 80.)

[Mairal 2008a] J. Mairal, M. Elad and G. Sapiro. *Sparse Representation for Color Image Restoration.* IEEE Transactions on Image Processing, vol. 17, no. 1, pages 53–69, Jan 2008. (Cited on pages 41 and 84.)

[Mairal 2008b] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro and Andrew Zisserman. *Supervised Dictionary Learning.* CoRR, vol. abs/0809.3083, 2008. (Cited on pages 41 and 45.)

[Mairal 2008c] Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert and Jean Ponce. *Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation.* In Proceedings of the 10th European Conference on Computer Vision: Part III, ECCV '08, pages 43–56, Berlin, Heidelberg, 2008. Springer-Verlag. (Cited on page 46.)

[Mairal 2009a] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman. *Non-local sparse models for image restoration.* In Computer Vision, 2009 IEEE 12th International Conference on, pages 2272–2279, Sept 2009. (Cited on page 43.)

[Mairal 2009b] Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro. *Online Dictionary Learning for Sparse Coding.* In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM. (Cited on pages 38, 50 and 88.)

[Mallat 1993] S. G. Mallat and Zhifeng Zhang. *Matching pursuits with time-frequency dictionaries.* IEEE Transactions on Signal Processing, vol. 41, no. 12, pages 3397–3415, Dec 1993. (Cited on pages 28 and 30.)

[Martínez 1998] Aleix Martínez and Robert Benavente. *The AR Face Database.* Technical report 24, Computer Vision Center, Bellatera, Jun 1998. Cites in Scholar Google: http://scholar.google.com/scholar?hl=en&lr=&client=firefox-a&cites=1504264687621469812. (Cited on page 45.)

[McDonald 2001] W. Ian McDonald, Alistair Compston, Gilles Edan, Donald Goodkin, Hans-Peter Hartung, Fred D. Lublin, Henry F. McFarland, Donald W. Paty, Chris H. Polman, Stephen C. Reingold, Magnhild Sandberg-Wollheim, William Sibley, Alan Thompson, Stanley Van Den Noort, Brian Y. Weinshenker and Jerry S. Wolinsky. *Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis.* Annals of Neurology, vol. 50, no. 1, pages 121–127, 2001. (Cited on page 78.)

[Miller 2004] D. H. Miller, M. Filippi, F. Fazekas, J. L. Frederiksen, P. M. Matthews, X. Montalban and C. H. Polman. *Role of magnetic resonance imaging within diagnostic criteria for multiple sclerosis.* Annals of Neurology, vol. 56, no. 2, pages 273–278, 2004. (Cited on page 76.)

[Mortazavi 2012] Daryoush Mortazavi, Abbas Z. Kouzani and Hamid Soltanian-Zadeh. *Segmentation of multiple sclerosis lesions in MR images: a review.* Neuroradiology, vol. 54, no. 4, pages 299–320, 2012. (Cited on page 81.)

[Nyul 2000] L. G. Nyul, J. K. Udupa and Xuan Zhang. *New variants of a method of MRI scale standardization.* IEEE Transactions on Medical Imaging, vol. 19, no. 2, pages 143–150, Feb 2000. (Cited on page 80.)

[Olshausen 1996] Bruno A. Olshausen and David J. Field. *Emergence of simple-cell receptive field properties by learning a sparse code for natural images.* Nature, vol. 381, pages 607 – 609, 1996. (Cited on page 43.)

[Olshausen 1997] Bruno A. Olshausen and David J. Field. *Sparse coding with an overcomplete basis set: A strategy employed by V1?* Vision Research, vol. 37, no. 23, pages 3311 – 3325, 1997. (Cited on page 43.)

[Palŭs 1992] Milan Palŭs and Ivan Dvořák. *Singular-value Decomposition in Attractor Reconstruction: Pitfalls and Precautions.* Phys. D, vol. 55, no. 1-2, pages 221–234, February 1992. (Cited on page 59.)

[Pati 1993] Y. C. Pati, R. Rezaiifar and P. S. Krishnaprasad. *Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition.* In Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on, volume 1, pages 40–44, Nov 1993. (Cited on page 31.)

[Peyré 2009] G. Peyré. *Sparse Modeling of Textures.* Journal of Mathematical Imaging and Vision, vol. 34, no. 1, pages 17–31, 2009. (Cited on page 84.)

[Pham 1999] D. L. Pham and J. L. Prince. *Adaptive fuzzy segmentation of magnetic resonance images.* IEEE Transactions on Medical Imaging, vol. 18, no. 9, pages 737–752, Sept 1999. (Cited on page 83.)

[Polman 2005] Chris H. Polman, Stephen C. Reingold, Gilles Edan, Massimo Filippi, Hans-Peter Hartung, Ludwig Kappos, Fred D. Lublin, Luanne M. Metz, Henry F. McFarland, Paul W. O'Connor, Magnhild Sandberg-Wollheim, Alan J. Thompson, Brian G. Weinshenker and Jerry S. Wolinsky. *Diagnostic criteria for multiple sclerosis: 2005 revisions to the ?McDonald Criteria?* Annals of Neurology, vol. 58, no. 6, pages 840–846, 2005. (Cited on page 78.)

[Polman 2011] Chris H. Polman, Stephen C. Reingold, Brenda Banwell, Michel Clanet, Jeffrey A. Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, Fred D. Lublin, Xavier Montalban, Paul O'Connor, Magnhild Sandberg-Wollheim, Alan J. Thompson, Emmanuelle Waubant, Brian Weinshenker and Jerry S. Wolinsky. *Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria.* Annals of Neurology, vol. 69, no. 2, pages 292–302, 2011. (Cited on page 78.)

[Prastawa 2008] M. Prastawa and G. Gerig. *Automatic MS Lesion Segmentation by Outlier Detection and Information Theoretic Region Partitioning*. 07 2008. (Cited on page 80.)

[Raina 2007] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer and Andrew Y. Ng. *Self-taught Learning: Transfer Learning from Unlabeled Data*. In Proceedings of the 24th International Conference on Machine Learning, ICML '07, pages 759–766, New York, NY, USA, 2007. ACM. (Cited on page 43.)

[Ramirez 2010] I. Ramirez, P. Sprechmann and G. Sapiro. *Classification and clustering via dictionary learning with structured incoherence and shared features*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 3501–3508, June 2010. (Cited on pages 39 and 45.)

[Ramirez 2012] I. Ramirez and G. Sapiro. *An MDL Framework for Sparse Coding and Dictionary Learning*. IEEE Transactions on Signal Processing, vol. 60, no. 6, pages 2913–2927, June 2012. (Cited on page 84.)

[Ren 2015] Huamin Ren, Weifeng Liu, Soren Ingvor Olsen, Sergio Escalera and Thomas B. Moeslund. *Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection*. In Xianghua Xie, Mark W. Jones and Gary K. L. Tam, editors, Proceedings of the British Machine Vision Conference (BMVC), pages 28.1–28.13. BMVA Press, September 2015. (Cited on page 45.)

[Roberts 2006] Lynne Roberts and Tom Miller. Beginner's guide to ms: Third edition. 2006. (Cited on page 75.)

[Rodriguez 2007] F. Rodriguez and G. Sapiro. *Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries*. IMA Preprint 2213, 2007. (Cited on page 46.)

[Rubinstein 2010a] R. Rubinstein, A. M. Bruckstein and M. Elad. *Dictionaries for Sparse Representation Modeling*. Proceedings of the IEEE, vol. 98, no. 6, pages 1045–1057, June 2010. (Cited on pages 33 and 37.)

[Rubinstein 2010b] R. Rubinstein, M. Zibulevsky and M. Elad. *Double Sparsity: Learning Sparse Dictionaries for Sparse Signal Approximation*. IEEE Transactions on Signal Processing, vol. 58, no. 3, pages 1553–1564, March 2010. (Cited on pages 37 and 84.)

[Scholkopf 1998] Bernhard Scholkopf, Alexander Smola, Er Smola and Klaus-Robert Müller. *Nonlinear Component Analysis as a Kernel Eigenvalue Problem*. Neural Computation, vol. 10, pages 1299–1319, 1998. (Cited on page 24.)

[Settles 2010] Burr Settles. *Active learning literature survey*. Technical report, 2010. (Cited on page 19.)

[Smith 2002] Stephen M. Smith. *Fast robust automated brain extraction*. Human Brain Mapping, vol. 17, no. 3, pages 143–155, 2002. (Cited on pages 80, 84 and 88.)

[Song 2012] Hyun Oh Song, Stefan Zickler, Tim Althoff, Ross Girshick, Mario Fritz, Christopher Geyer, Pedro Felzenszwalb and Trevor Darrell. Computer vision – eccv 2012: 12th european conference on computer vision, florence, italy, october 7-13, 2012, proceedings, part ii, chapter Sparselet Models for Efficient Multiclass Object Detection, pages 802–815. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. (Cited on page 43.)

[Souplet 2008] J. Souplet, C. Lebrun, N. Ayache and G. Malandain. *An Automatic Segmentation of T2-FLAIR Multiple Sclerosis Lesions*. 07 2008. (Cited on page 83.)

[Sweeney 2013] Elizabeth M. Sweeney, Russell T. Shinohara, Navid Shiee, Farrah J. Mateen, Avni A. Chudgar, Jennifer L. Cuzzocreo, Peter A. Calabresi, Dzung L. Pham, Daniel S. Reich and Ciprian M. Crainiceanu. *{OASIS} is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in {MRI}*. NeuroImage: Clinical, vol. 2, pages 402 – 413, 2013. (Cited on page 82.)

[Tibshirani 1994] Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso*. Journal of the Royal Statistical Society, Series B, vol. 58, pages 267–288, 1994. (Cited on pages 32 and 46.)

[Tintoré 2000] Mar Tintoré, Alex Rovira, Maria J. Martínez, Jordi Rio, Pablo Díaz-Villoslada, Luis Brieva, Cecilia Borrás, Elisenda Grivé, Jaume Capellades and Xavier Montalban. *Isolated Demyelinating Syndromes: Comparison of Different MR Imaging Criteria to Predict Conversion to Clinically Definite Multiple Sclerosis*. American Journal of Neuroradiology, vol. 21, no. 4, pages 702–706, 2000. (Cited on page 79.)

[Tong 2013] Tong Tong, Robin Wolz, Pierrick Coupe, Joseph V. Hajnal and Daniel Rueckert. *Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling.* NeuroImage, vol. 76, no. 0, pages 11 – 23, 2013. (Cited on page 84.)

[Tustison 2010] N.J. Tustison, B.B. Avants, P.A. Cook, Yuanjie Zheng, A. Egan, P.A. Yushkevich and J.C. Gee. *N4ITK: Improved N3 Bias Correction.* IEEE Transactions on Medical Imaging, vol. 29, no. 6, pages 1310–1320, June 2010. (Cited on pages 83 and 88.)

[Vovk 2007] U. Vovk, F. Pernus and B. Likar. *A Review of Methods for Correction of Intensity Inhomogeneity in MRI.* IEEE Transactions on Medical Imaging, vol. 26, no. 3, pages 405–421, March 2007. (Cited on page 80.)

[Wang 2012] Yun-Heng Wang, Jun-Bao Li and Ping Fu. *Medical Image Super-resolution Analysis with Sparse Representation.* In Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pages 106–109, July 2012. (Cited on page 84.)

[Weiss 2013] Nick Weiss, Daniel Rueckert and Anil Rao. *Multiple Sclerosis Lesion Segmentation Using Dictionary Learning and Sparse Coding.* In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013, volume 8149, pages 735–742. 2013. (Cited on pages 45, 80, 84, 87 and 102.)

[Wells 1996a] W. M. Wells, W. E. L. Grimson, R. Kikinis and F. A. Jolesz. *Adaptive segmentation of MRI data.* IEEE Transactions on Medical Imaging, vol. 15, no. 4, pages 429–442, Aug 1996. (Cited on page 83.)

[Wells 1996b] William Wells, Paul Viola, Hideki Atsumi, Shin Nakajima and Ron Kikinis. *Multi-modal volume registration by maximization of mutual information.* Medical Image Analysis, vol. 1, no. 1, pages 35 – 51, 1996. (Cited on pages 83 and 88.)

[Wright 2009] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma. *Robust Face Recognition via Sparse Representation.* IEEE Trans. on PAMI, vol. 31, no. 2, pages 210–227, Feb 2009. (Cited on pages 38, 41 and 84.)

[Yang 2010a] J. Yang, J. Wright, T. S. Huang and Y. Ma. *Image Super-Resolution Via Sparse Representation.* IEEE Transactions on Image

Processing, vol. 19, no. 11, pages 2861–2873, Nov 2010. (Cited on pages 41 and 43.)

[Yang 2010b] Meng Yang, Lei Zhang, Jian Yang and Dejing Zhang. *Metaface learning for sparse representation based face recognition.* In Image Processing (ICIP), 2010 17th IEEE International Conference on, pages 1601–1604. IEEE, 2010. (Cited on pages 39 and 45.)

[Yang 2011] Meng Yang, D. Zhang, Xiangchu Feng and D. Zhang. *Fisher Discrimination Dictionary Learning for sparse representation.* In IEEE ICCV, pages 543–550, Nov 2011. (Cited on pages 40, 45, 48, 68 and 69.)

[Yu 2013] Nan-Nan Yu, Tian-Shuang Qiu and Wen-hong Liu. *Medical Image Fusion Based on Sparse Representation with KSVD.* In Mian Long, editor, World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China, volume 39 of *IFMBE Proceedings*, pages 550–553. Springer Berlin Heidelberg, 2013. (Cited on page 84.)

[Zhang 2000] G. P. Zhang. *Neural networks for classification: a survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 30, no. 4, pages 451–462, Nov 2000. (Cited on page 17.)

[Zhang 2009] Wei Zhang, Akshat Surve, Xiaoli Fern and Thomas Dietterich. *Learning Non-redundant Codebooks for Classifying Complex Objects.* In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 1241–1248, New York, NY, USA, 2009. ACM. (Cited on page 46.)

[Zhang 2010a] Qiang Zhang and Baoxin Li. *Discriminative K-SVD for dictionary learning in face recognition.* In IEEE CVPR, pages 2691–2698, June 2010. (Cited on pages 40, 41 and 45.)

[Zhang 2010b] Shaoting Zhang, Junzhou Huang, D. Metaxas, Wei Wang and Xiaolei Huang. *Discriminative sparse representations for cervigram image segmentation.* In IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pages 133–136, April 2010. (Cited on page 84.)

[Zhang 2012a] Shaoting Zhang, Yiqiang Zhan, Maneesh Dewan, Junzhou Huang, Dimitris N. Metaxas and Xiang Sean Zhou. *Towards robust and effective shape modeling: Sparse shape composition.* Medical Image Analysis, vol. 16, no. 1, pages 265 – 277, 2012. (Cited on page 84.)

[Zhang 2012b] Shaoting Zhang, Yiqiang Zhan and Dimitris N. Metaxas. *Deformable segmentation via sparse representation and dictionary learning.* Medical Image Analysis, vol. 16, no. 7, pages 1385 – 1396, 2012. Special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention. (Cited on page 84.)

[Zhu 2005] Xiaojin Zhu. *Semi-supervised learning literature survey.* 2005. (Cited on page 19.)

[Zijdenbos 2002] A.P. Zijdenbos, R. Forghani and A.C. Evans. *Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis.* IEEE Transactions on Medical Imaging, vol. 21, no. 10, pages 1280–1291, Oct 2002. (Cited on page 98.)

Résumé: La plupart des signaux naturels peuvent être représentés par une combinaison linéaire de quelques atomes dans un dictionnaire. Ces représentations parcimonieuses et les méthodes d'apprentissage de dictionnaires (AD) ont suscité un vif intérêt au cours des dernières années. Bien que les méthodes d'AD classiques soient efficaces dans des applications telles que le débruitage d'images, plusieurs méthodes d'AD discriminatifs ont été proposées pour obtenir des dictionnaires mieux adaptés à la classification. Dans ce travail, nous avons montré que la taille des dictionnaires de chaque classe est un facteur crucial dans les applications de reconnaissance des formes lorsqu'il existe des différences de variabilité entre les classes, à la fois dans le cas des dictionnaires classiques et des dictionnaires discriminatifs. Nous avons validé la proposition d'utiliser différentes tailles de dictionnaires, dans une application de vision par ordinateur, la détection des lèvres dans des images de visages, ainsi que par une application médicale plus complexe, la classification des lésions de scléroses en plaques (SEP) dans des images IRM multimodales. Les dictionnaires spécifiques à chaque classe sont appris pour les lésions et les tissus cérébraux sains. La taille du dictionnaire pour chaque classe est adaptée en fonction de la complexité des données. L'algorithme est validé à l'aide de 52 séquences IRM multimodales de 13 patients atteints de SEP.
Mot clés: Représentations parcimonieuses, apprentissage, SEP, IRM.

Abstract: Most natural signals can be approximated by a linear combination of a few atoms in a dictionary. Such sparse representations of signals and dictionary learning (DL) methods have received a special attention over the past few years. While standard DL approaches are effective in applications such as image denoising or compression, several discriminative DL methods have been proposed to achieve better image classification. In this thesis, we have shown that the dictionary size for each class is an important factor in the pattern recognition applications where there exist variability difference between classes, in the case of both the standard and discriminative DL methods. We validated the proposition of using different dictionary size based on complexity of the class data in a computer vision application such as lips detection in face images, followed by more complex medical imaging application such as classification of multiple sclerosis (MS) lesions using MR images. The class specific dictionaries are learned for the lesions and individual healthy brain tissues, and the size of the dictionary for each class is adapted according to the complexity of the underlying data. The algorithm is validated using 52 multi-sequence MR images acquired from 13 MS patients.
Keywords: Sparse representations, machine learning, multiple sclerosis, MRI.