Visualizing Large Medical Image Datasets using the 3D Scale-Invariant Feature Transform (SIFT)

by

Marzieh ROKOOIE

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN ELECTRICAL ENGINEERING
M.A.Sc.

MONTREAL, APRIL 8, 2020

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

M. Matthew Toews, Thesis Supervisor
Department of Automated Production Engineering, École de technologie supérieure

M. Stephane Coulomb, President of the Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Mrs. Silvie Ratté, External Examiner
Department of Software Engineering and IT, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MARCH 19, 2020

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Matthew Toews, for his useful comments, notices and consultation through the learning process of this master thesis. Furthermore, I would like to show my gratefulness to Etienne Pepin for his great contribution in my work. To my beloved family, thank for your endless love, support, sacrifice, guidance, and patience since I was born. I would like to give thanks my loved ones, who supported me for the duration of entire process, both by keeping me motivated and helping me putting pieces together. I will be grateful forever for your love.

Visualisation de grands ensembles d'images médicales utilisation de la transformation de caractéristique invariante à l'échelle 3D (SIFT)

Marzieh ROKOOIE

RÉSUMÉ

Ce mémoire propose un paradigme de classification des images dans lequel, au lieu d'une classification entièrement automatique, l'objectif est de générer un résumé visuel des informations d'image relatives aux classes. Audela de fornir un simple classification, ex. malade ou en santé dans le contexte d'imagérie médicale, notre approche fournit une visualisation des informations liées à la classification. Dans cette recherche particulière, nous fournissons une étude de la classification et de la visualisation basées sur les caractéristiques locales 3D « Scale-Invariant Transform » (SIFT). Nous avons développé un cadre probabiliste bayésien permettant de classer une nouvelle image et de visaliser la probabilité de la classe étant donné l'endroit spatiale dans l'image. Un nouvel estimateur de la variance est proposé basé sur le test exact de Fisher. Nous avons utilisé l'implementation SIFT-Rank Toews & Wells (2013) pour l'extraction des caractéristiques 3D SIFT et implémenté les techniques de classification et visualisation en code MATLAB. Nous avons démontré l'utilité de notre méthode avec une analyse des images de résonance magnétiques (IRM) des cerveaux humains adults à partir des données « Open Access Series of Imaging Studies » (OASIS) Marcus *et al.* (2007), sur trois catégories binaires: âge (jeune, vieux), sexe (mâle femelle) et maladie (maladie d'Alzheimer, contrôles normaux). La plus discriminante de ses catégories fut celle de l'age avec une taux de classification de 94.71%. La méthode présenté pourrait donc trouver application dans la suivi du vieillissement du cerveau, et pourrait s'appliquer de façon générale aux autres modalités d'images et régions d'intéret, ex. imagerie par tomodensitométrie (TDM) du pumon.

Mots-clés: Visualisation d'Images Médicales, Apprentissage Automatique, Classification, Données Massives, Images Cérébrales

Visualizing Large Medical Image Datasets
using the 3D Scale-Invariant Feature Transform (SIFT)

Marzieh ROKOOIE

ABSTRACT

This thesis proposes an image classification paradigm where instead of fully automatic classification, the goal is to generate a highly-informative visual summary of class-related information for human interpretation. Rather than providing a single classification, we provide a visualisation highlighting the information relevant to group differences. In this particular research, we provide a survey of instance-based classification and visualization. A probabilistic framework is developed for classification and visualization, based on the 3D scale-invariant feature transform (SIFT) format. We propose a novel kernel density bandwidth estimator for SIFT feature densities, based on hypothesis testing, where the bandwidth minimizes the p-value of Fisher's exact test. We also propose a method of of ranking features based on the false discovery rate (FDR). An existing implementation of the SIFT-Rank method Toews & Wells (2013) is used for feature extraction, and classification and visualization are implemented in MATLAB. We validate our approach on 3D magnetic resonance image (MRI) data of the adult human brain from the Open Access Series of Imaging Studies (OASIS) dataset Marcus *et al.* (2007). Experiments investigated classification and visualisation of three binary categories: age (young, old), gender (male, female), and disease (Alzheimer's disease vs. healthy). The highest classification accuracy was 94.71% for age (old vs. young), and the method may prove useful for understanding the aging process. The method is generally applicable to arbitrary 3D medical image modalities and conditions, for example computed tomography (CT) lung scans.

Keywords: Visualization, Medical Image Data, Machine Learning, Classification, Big Data, Human Brain, Magnetic Resonance Image, Scale-invariant Feature Transform

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABREVIATIONS

ÉTS	École de Technologie Supérieure
OASIS	Open Access Series of Imaging Studies
MRI	Magnetic Resonance Imaging
MR	Magnetic Resonance
kNN	k-Nearest Neighbors
SIFT	Scale-Invariant Feature Transform
KDE	Kernel Density Estimation
2D	Two Dimensional
RoI	Regions of Interest
MAP	Maximum a-Posterior
DoG	Difference-of-Gaussian
GoH	Gradient Orientation Histogram
IID	Independent, Identically Distributed
CNN	Convolutional Neural Networks
LF-Net	Local Feature Network
LIFT	Learned Invariant Feature Transform
DNNs	Deep Neural Networks
3D	Three Dimensional
SVM	Support Vector Machines

CDR	Clinical Dementia Rating
FDS	Feature-Distance Space

INTRODUCTION

Consider human-in-the-loop systems that rely on a human expert to make decisions based on data, e.g., clinicians such as radiologists or pathologists interpreting images visually to identify the subtle signs of disease to plan the best possible treatment. According to this fact, one of the most challenging fields is medical image processing these days.

Rapidly interpreting large numbers of images is a tedious and time-consuming task which is accomplished by radiologists or clinical experts. The classification result's accuracy depends on the expert human's experience and is prone to human error. Automatic classification is a powerful tool; however even modern approaches may produce erroneous results with high confidence, Nguyen *et al.* (2015), and it may be challenging to provide an optimal decision from a single classification label without the rich-nuanced information presented in the original image.

According to the age of computers and improvements in graphical user interfaces, the definition of the word "visualization" has changed over the time, from a form of cognitive/intellectual sketch of something to a graphical illustration of an object or information set and also cause to the rise of exploratory data analysis. Data analysis has historically been a statistical issue while many common types of visualizations like scatter plots or box plots originate from statistics, Friendly & Denis (2001).

What is the best information code to provide to the human expert to balance between fully manual and fully automatic approaches. We propose providing the human visual system with a visualization that simultaneously is highly informative regarding diagnosis/classification label and contains spatially pertinent information indicating how the diagnosis/classification is obtained.

In our study, the principal contribution is to visualize significance features which aims to automatically identify subject's age range, gender, and also discriminate between healthy subjects and diseased subjects with Alzheimer, by their whole-brain magnetic resonance images, MRI. MRI is a popular high-quality medical imaging technique without any bad radiation effect on the body. MRI is the best choice to distinguish tissue characteristic differences easily. Soft tissue contrast and being non-invasive are the clearest advantages of MR method, Haacke *et al.* (1999).

In this work, to have an automatic classification, a nonlinear classifier and a statistical significance test are used to detect the most significant features of images in group-differences of visualization/classification challenge. k-Nearest Neighbors-kernel density-based algorithm (kNN-kernel density-based), a non-parametric classification/regression method in pattern recognition is used as automatic classifier technique, Altman (1992); Cover & Hart (1967); Tran *et al.* (2006). k refers to the closest training samples which are considered as the input of the classifier. Fisher's exact test which is a practical method in categorical data test is applied to find the feature's informative closest samples. The significance of the association/contingency is examined between the results of the classifications, Fisher (1922).

Figure 0.1 illustrates the main objective of this work which is to visualize the region of interests of the test subject which are found based on the k nearest neighbors of the training dataset.

We, therefore, begin with the review of related literature in chapter 1. We have a brief look at the history of feature extraction. We show why local features are more important rather than global ones in this work, how the scale-invariant feature transform, SIFT, is more reliable than other techniques. We show how automatic image classification techniques have progressed rapidly these years. And the primary goal of this chapter is to show how kernel density classification was developed using the statistical significance test to estimate the kernel parameters.

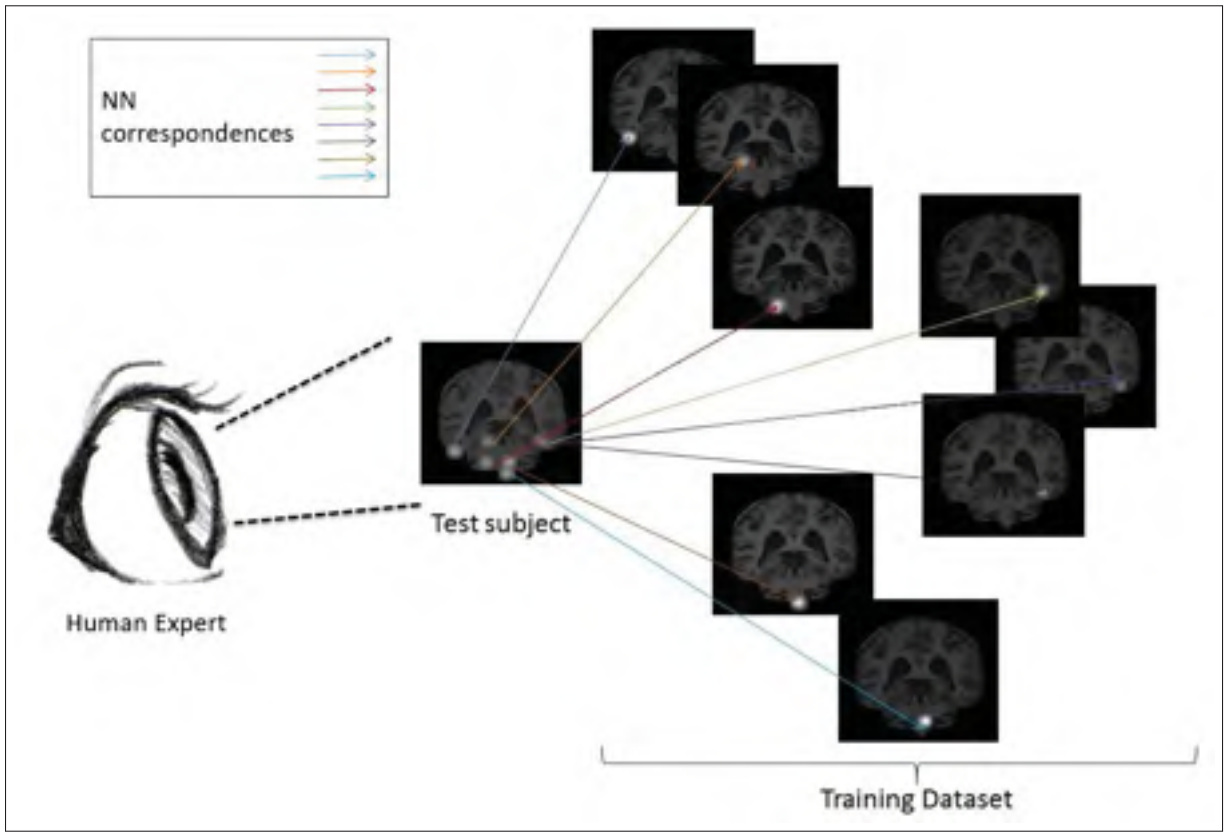

Figure 0.1 A human expert can view and interpret an MRI of a test subject in terms of RoIs identified by k-NN correspondences to a training database

In chapter 2 we present the general probabilistic model and the mathematics which are used in this work. All the contributions are listed in this chapter as follow: 2.1 Local Features, 2.2 Generative Model, 2.3 Distance Metrics, 2.4 Classification, 2.5 Parameter Estimation and 2.6 Visualization.

In Chapter 3, we present experimentation to validate our proposed method. There are many comparisons between different datasets results. In general, the trials show this method is useful in automatic classification, and visualization aims to distinguish the regions of interest. This work might be helpful in a wide variety of medical contexts.

We conclude the work with a discussion and offer pointers to future work.

CHAPTER 1

RELATED WORKS

This study presents a general model to generate a highly-informative visual summary of class-related information for human interpretation, based on 3D scale-invariant image features. Here, we review prior works on the local image features, classification, and visualization.

1.1 Local Image Features

An image typically contains an enormous amount of information illustrated by an array or lattice of intensity measurements in computer vision programs, i.e., pixels in 2D photographs or voxels in 3D MRI volumes. In images of natural objects or scenes, most of the images may be composed of redundant or uninformative intensity information, for example, regions of homogeneous image intensity. Information tends to be concentrated into a small subset of unique or distinctive regions, i.e., features. Features may contain global attributes information of the image, e.g., intensity histogram, frequency domain descriptors, covariance matrix, and high order statistics, etc., or include local region information of the image, e.g., spatially localized edges, corners or blob patterns, etc. The image can be defined as a set of such global or local regions, Figure 1.1, which are known in the computer vision literature as global image and local image features, respectively. As it is shown by Figure 1.1, global features contain overall information such as shape, whereas local features focus on the details. In this work, we discuss the anatomical structures in different subjects images, and we seek to observe and characterize similarities and differences between them by detail. Therefore local features are more effective rather than global.

Local features are considered for essential reasons. There is data reduction compared to the original image information. As the amount of information reduced by finding local features, the processing time of algorithm would be optimized and considerably faster. Furthermore, spatially localized features are robust to the occlusion and clutter, rotation, translation and resolution, Lowe (2004); Wells III (1997); Fergus *et al.* (2007). These interesting points could be matched

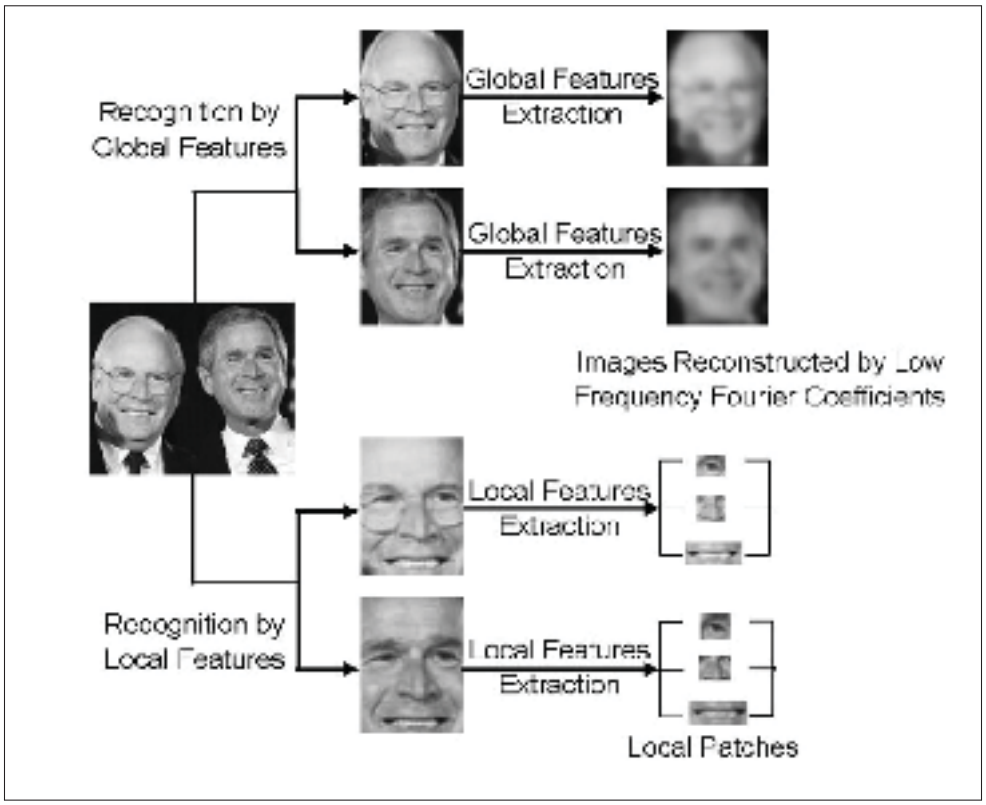

Figure 1.1 Global and local image features in the context of the face recognition task
Taken from Su *et al.* (2009)

across images without an explicit search or image registration, Amit & Kong (1996); Yang *et al.* (2011). Moreover, local similarities detection between images is more reliable rather than global similarities, Toews *et al.* (2010); Toews & Arbel (2009).

Local features are useful in many applications e.g., image alignment, reconstruction, motion tracking, object recognition, indexing and database retrieval, navigation, etc. In fact, the local feature representation might be considered as a general building block for many computer vision and medical imaging algorithms.

Local interest points detectors were used to identify the salient points in order to match the images in binocular vision by Marr & Poggio (1977) and robotic mapping by Moravec (1979) in early works. Later Harris & Stephens (1988) and Rohr (1997) detected the corner and landmark by calculating spatial gradients. This was generally achieved via salience operators evaluating

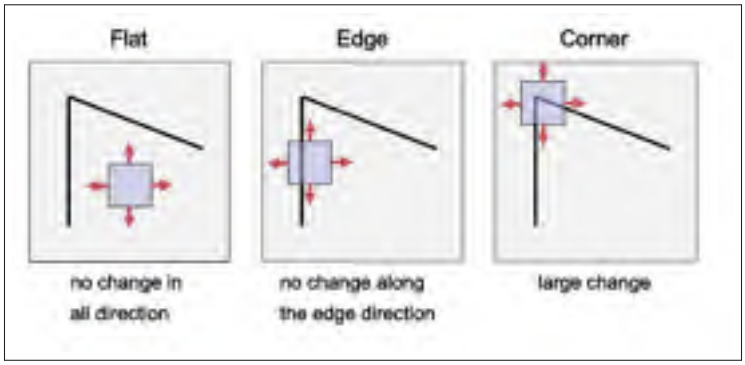

Figure 1.2 Harris Corner Detection, the corner has been detected if a significant change in appearance occurred by shifting the window in any directions Taken from Harris & Stephens (1988)

fixed-size image regions. Figure 1.2 illustrates a combined corner and edge detector technique which is done by Harris & Stephens (1988). The shift-able window evaluates the intensity changes by shifting in any directions.

Harris corner detector's basic formula is shown by Equation 1.1, Harris & Stephens (1988). This is one of the early attempts to find the corners. The change intensity is detected by displacing the window in all directions based on Taylor Series:

$$E(u,v) = \sum_{x,y} \underbrace{w(x,y)}_{\text{window function}} \; [\underbrace{I(x+u,y+v)}_{\text{shifted intensity}} - \underbrace{I(x,y)}_{\text{intensity}}]^2 \tag{1.1}$$

This window function points to the rectangular window or Gaussian window with the center location of (x, y).In Eq. 1.1, term I refers to the pixel intensity and (u, v) shows the window location displacement. This function gives weights to the pixels underneath. The maximization of $E(u, v)$ may be corresponded to the corners or edges. Eq. 1.1 can be rewritten as a matrix formed equation and we may have:

$$E(u,v) \approx (u,v) M \begin{pmatrix} x \\ y \end{pmatrix} \tag{1.2}$$

where M is :

$$M = w(x, y) \begin{pmatrix} \sum_{(x,y)} I_x^2 & \sum_{(x,y)} I_x I_y \\ \sum_{(x,y)} I_x I_y & \sum_{(x,y)} I_y^2 \end{pmatrix} \tag{1.3}$$

Pennec *et al.* (2000) and Jian & Vemuri (2011) showed that we may use local feature geometry to describe the image pattern independent of its density, e.g. through landmarks or point sets. Shi *et al.* (1994), Rohr *et al.* (2001), and Urschler *et al.* (2006) presented procedures in order to have impressive image corresponding by encoding image intensity information related to the features.

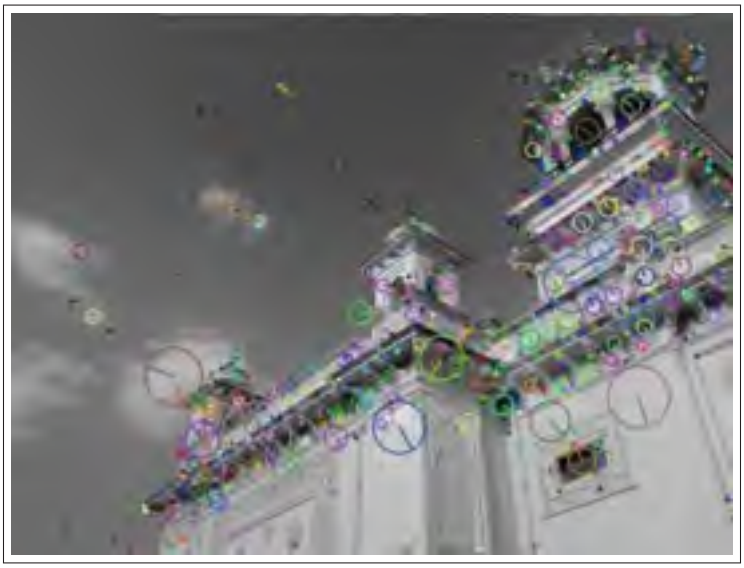

Figure 1.3 Illustration of distinct structures which are identified by local features
Taken from Abid (2013)

An important development was scale invariant feature detection, where scale-space theory proposed by Lindeberg (1998) and Romeny (2008) was used to identify the invariant features to the size or scale of image patterns in addition to the location. Therefore, interest operators are extended in this framework. This idea reflects the image pattern and the scale of patterns in the observed images are linked.

Lowe (2004) introduced the local invariant image features e.g. the scale-invariant feature transform (SIFT) method, which is one of the most famous algorithm in image matching in

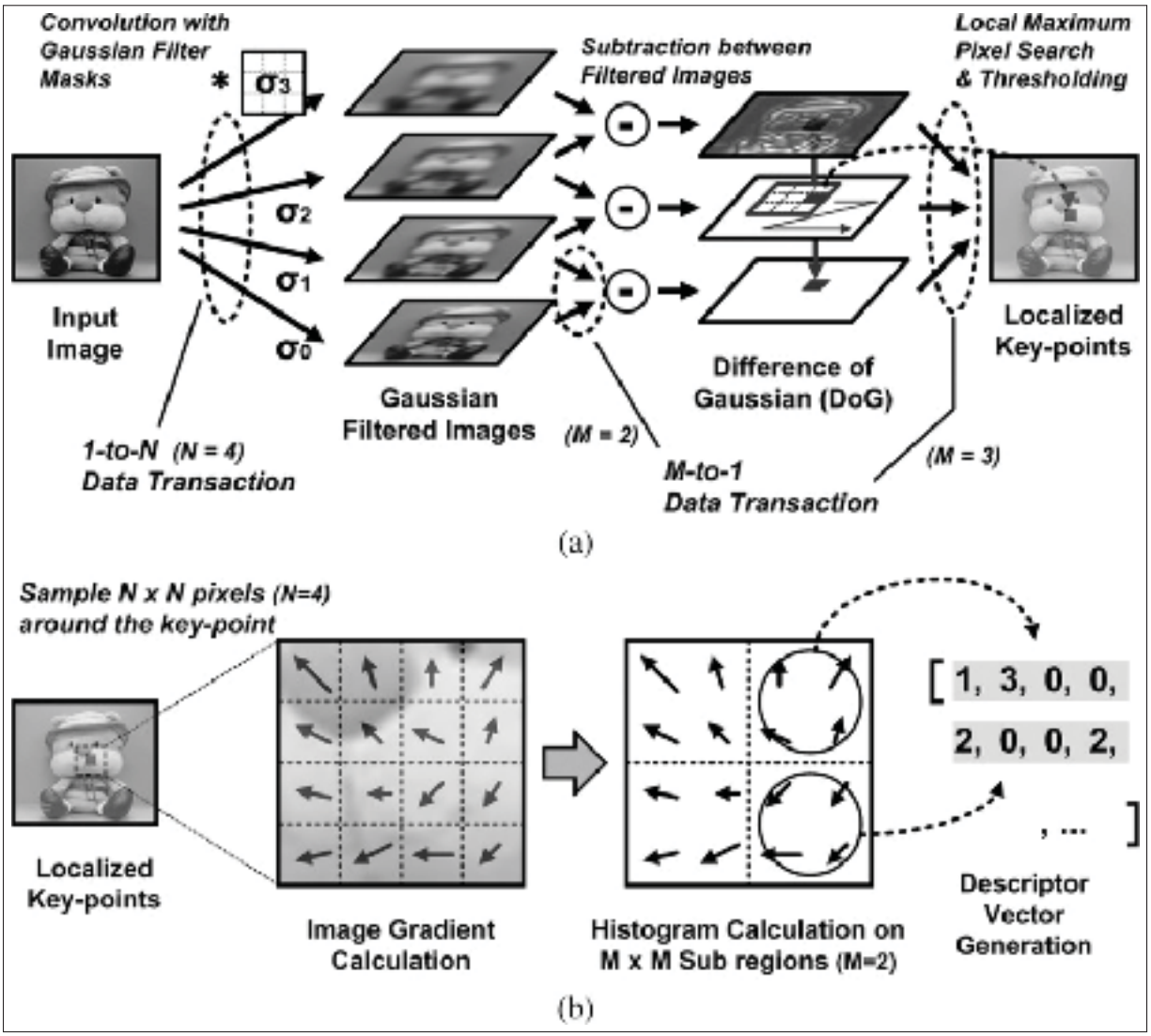

Figure 1.4 The overall flow of the SIFT computation which is divided into (a) key-point localization and (b) descriptor vector generation stages
Taken from Kim *et al.* (2009)

the computer vision community, Mikolajczyk & Schmid (2004). The corresponding global geometrical similarities might have been computed by the Scale-invariant features which can be extracted over and over. As well they can be used for affine deformations Mikolajczyk *et al.* (2005). The methodology of scale-invariant feature extraction method is basically a search on image regions to detect the maximal of a criterion of salience, e.g. the magnitude of Gaussian derivatives in scale Lowe (2004) and/or space Mikolajczyk & Schmid (2004), image phase

Carneiro & Jepson (2003) or information-theoretic measures such as entropy Kadir & Brady (2001) or mutual information Toews & Wells (2010). Fig1.4 shows the overall flow of the SIFT computation.

In general, local feature methods have been operated by detecting salient or interesting image regions then encoding or describing these regions for image-to-image correspondence. Invariant detection involves repeatably detecting the same image regions despite geometrical changes, e.g. translation, scaling, rotation and possibly affine image deformations, Ke & Sukthankar (2004).

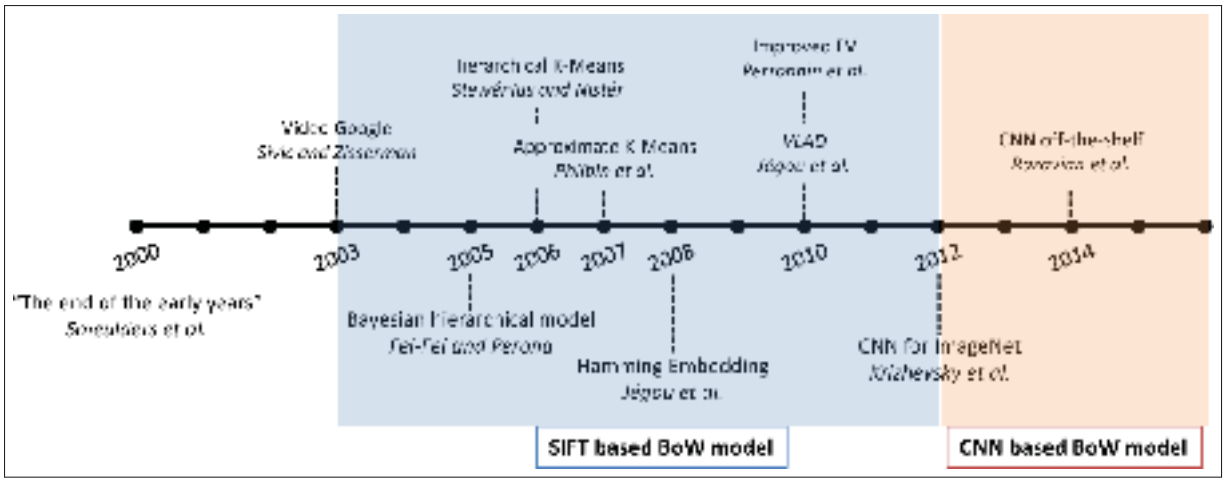

Figure 1.5 A quick overview of instance retrieval history. Following the pioneering work of convolutional neural network, CNN-based methods began to gradually take over, however SIFT-based methods were still moving forward
Taken from Zheng *et al.* (2017), pioneering work is done by Krizhevsky *et al.* (2012)

Local features have played a distinctive role in computer vision, becoming a standard for image matching, Hartley & Zisserman (2003). This method have been used despite of deep network alternatives. The state of the art methods typically include dense matching but they suffer from occlusions, which local features are robust against, Choy *et al.* (2016). Figure 1.5 illustrates a brief history of instance retrieval which utilize SIFT and also CNN over the time.

Figure 1.6 shows an updated history which is done by Zheng *et al.* (2017). In the updated version, there is more details after 2014 about state of the arts methods.

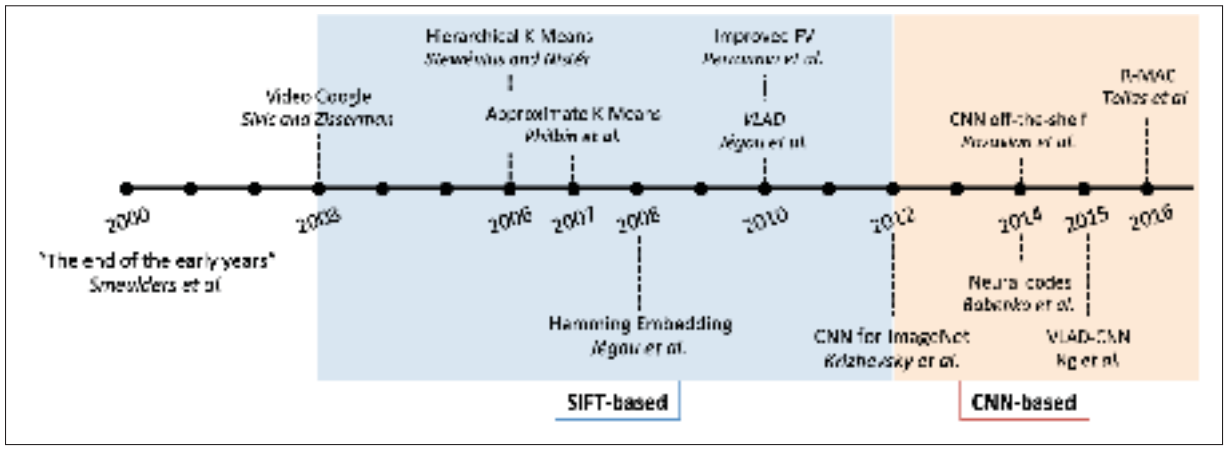

Figure 1.6 Updated overview of instance retrieval history. In 2014, a hybrid method is proposed which is extracting multiple CNN features from an image and fine-tuning a CNN model for generic instance retrieval was done for the first time
Taken from Zheng *et al.* (2017)

Figure 1.7 shows that how the state of the art methods such as CNN play a weak role for instance retrieval applications in comparison to the SIFT-based method under small-sized codebooks, Zheng *et al.* (2017). The updated version of this diagram is illustrated by figure 1.8.

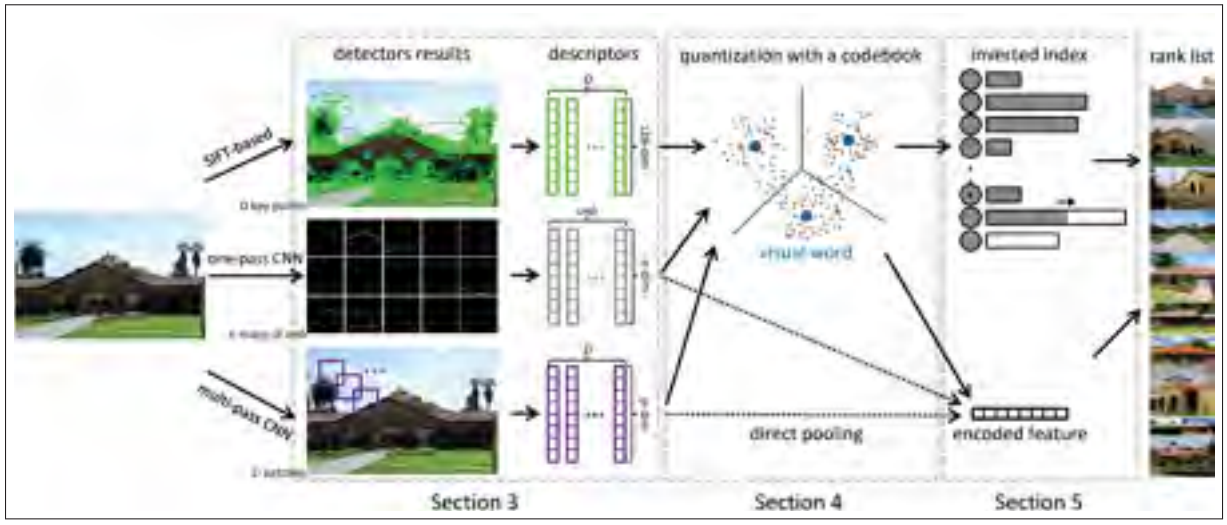

Figure 1.7 SIFT vs. CNN-based retrieval models. The inverted index is necessary be under large/mid-sized codebooks
Taken from Zheng *et al.* (2017)

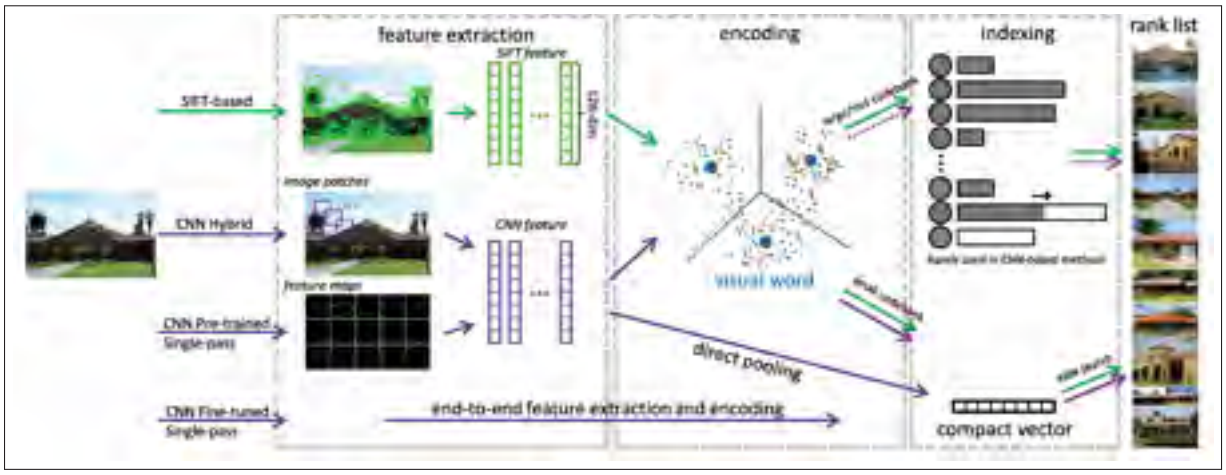

Figure 1.8 Updated SIFT vs. CNN-based retrieval models
Taken from Zheng *et al.* (2017)

Yi *et al.* (2016) have offered a novel Deep Network architecture, learned invariant feature transform, LIFT, which applies the full feature point handling structure such as detection, orientation, and feature description. However, previous works have successfully implemented each one of these problems individually, they use all three in a unified manner while preserving end-to-end differentiability. Laguna *et al.* (2019) have proposed a novel local feature detection approach that utilizes both handcrafted and learned CNN filters. Ono *et al.* (2018) have introduced a local feature network, LF-Net, a novel deep architecture to learn local features without hand-crafted features, i.e., SIFT. It includes the entire feature extraction pipeline, and can be trained end-to-end with just a collection of images.

1.2 Automatic Image Classification

Automatic classification is a computational task where the goal is to assign a label to a data sample. Fully automatic image classifiers trained via machine learning have long promised to alleviate this workload. However, there are challenges. First, it may be difficult to interpret the result of the classification. For example, modern classifiers such as deep convolutional neural networks (CNNs) achieve impressive classification results which can produce confidence values. However, they also produce highly confident classifications in case of irrelevant, artificially

generated images, Nguyen *et al.* (2015); Szegedy *et al.* (2013), raising the possibility of patient misclassification. Second, it is difficult to ensure classifiers generalize across imaging conditions due to limited numbers of training data. For example, MRI are notoriously difficult to normalize across the different sites. Accurate site-wise classifiers may be possible via combinations of transfer learning, classifier retraining. However, retraining is computationally intensive, requires many training samples (e.g patient data) that might not be available or easily accessible. Effective transfer learning for multi-site medical imaging data remains an open research topic.

Recently, there has been a drastic changes in the field of information technology and the worldwide web access to the visualized data. The main challenge in this field is organizing and classifying these data to make the user's access easy for appropriate data. Users wish to get an appropriate image when they search and also are interested in navigating through the images. These types of requirement has generated excessive demands for operational and flexible systems for organizing digital images and visual data.

One important method for solving such a problem is using image classification to constitute a digital library, Haralick *et al.* (1973). Image classification is the mission of splitting images into categories based on the labels which are presented in training data. There are various methods for image classification but a general issue is involved by this matter can be listed as follow:

- Image features; finding significant part of the image and express image by them,
- Organization of feature data; categorizing these features in a way to be identified separable,
- Classifier; dividing images in different categories.

Image classification could give rise to a semantic organization of a digital database. In order to map the irrelevant visual features it is important to train a large number of classifiers to accomplish large-scale image classification. The classifiers performance largely depends on the devices design for training and the quality of feature objectives which could be two serious matter. Different model of classifiers has been used recently. Vapnik (2013) explain about different classifiers and their structure. Mika *et al.* (1999)proposes an example of a non linear

classifier based on based on Fisher's discriminant. Too many challenging problems can be indicated for classification of large number of classes.

The inter-concept accordance, feature extraction method and the classification error for the relevant object classes are the number of these issues. To overcome these problems and have an appropriate image classification, there are several type of organization like hierarchical approach such as the layers in deep neural networks, DNNs introduced by Bengio *et al.* (2009), traditional approach, inter-related approach and kernel density and function, Girshick *et al.* (2014); Póczos *et al.* (2012).

The first modern convnet, which had some of the essential ingredients we still use in CNNs, were introduced by LeCun *et al.* (1989b) who applied a back-propagation learning algorithm to the convolutional neural network architecture implemented by Fukushima (1988).

In this work, we adopt an instance-based learning framework in which all data is maintained in memory. This has the advantage of allowing arbitrary queries based on all data. Representations derived convolutional networks must be trained for specific conditions, and may each new patient scan is compared to all previous patient data, and informative disease-related image regions are highlighted. In this project, we deal with many problems such as, large dimensions, large quantities, noise, and structural.

1.2.1 Kernel Density Estimation

A non-parametric method to calculate the probability density function is named kernel density estimation, KDE which is one of the basic method for smoothing data where a population is derived according to a finite data sample. It also can be used for classification if the asymptotic properties of non-parametric variable has been considered, Silverman (2018).

The kernel density estimation is introduced to take advantage of the density classification and the informative thresholds. Estimating a normalized probability density function from a set of sample training data points could be provided by KDE. KDE can identify an accurate distribution

by approximating most well-behaved arbitrary distributions with continuous second derivative in compare to parametric methods. For example histograms method need asymptotically more data to obtain the same error or mixture model of five Gaussian unable to accurately capture distributions that contain more than five distinct regions of high density,Sheather & Jones (1991).

The first approach for classification via non parametric density estimation was reported by Van Ryzin (1966) and Wolverton & Wagner (1969). Krzyzak & Pawlak (1984) study Classification procedures using variable kernel density estimation. They introduced classification procedures by variable kernel density and gave sufficient conditions for weak and strong stability of the estimation.

A learning algorithm structure has been introduced by Mei *et al.* (2014), which can provide a superior performance for effective training of large numbers of correlated classifiers in order to classifying many images with annotation. To find the similarity in visual context and construct a visual concept network, they use a kernel function and a visual concept network for determining the inter-concept visual inter-related directly and characterizing the correlations in learning tasks intuitively in the visual feature space rather than in the label space. The algorithm also yield an efficient result on large-scale image classification tasks.

Gan & Bailis (2017) reported a research on Scalable Kernel Density Classification via Threshold-Based Pruning. They could boost the efficiency of KDE to classify points by their density (density classification) by introducing a simple technique. A threshold kernel density classification with asymptotic speedups achievement by maintaining accuracy guarantees introduced as a new technique. They also found out that in density classification method application, much of the computational overhead in computing kernel density estimates is unnecessary.

Kernel methods can be used to estimate the subject's class related in this experiments. Kernels are non-linear functions defined over the features $x \in X$, Bishop *et al.* (1995) that may be expressed as general function of $k(x, x')$. Gaussian kernel function is expressed as follow:

$$k(x, x') = exp(-\frac{||x - x'||^2}{2\sigma^2}) \tag{1.4}$$

where $x, x' \in X$ and σ is standard deviation of the distribution. Kernels can be seen as generalized covariances. If we assume that x has an implicit representation of features, e.g. $\phi(x) \in F$, a kernel function can be seen as a dot product in this feature space.

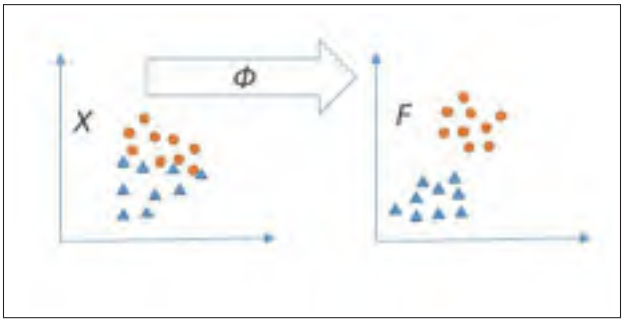

Figure 1.9 Feature representation by using kernel function

1.2.2 K-nearest Neighbors

Cover & Hart (1967) show that as the amount of data becomes very large, nearest neighbor methods approach a small multiple of Bayes optimal error. kNN is one of the non-parametric method used for pattern recognition. In this method, the k closest training examples are assumed as the input, and the class membership is the output. The distances and neighborhood is considered in the feature space. The greater number of votes of the neighbors can be classified the object. In other word, the object would be assigned to the class which is most frequented among the k-nearest neighbors. In the analysis of the probability tables, one of the common statistical test in order to find significance items is Fisher's exact test. It is considered as a component in the class of exact tests.

1.3 Visualization

Due to the growth of generated data, there are challenges for data analysis and interpretation. Data visualization aids to hand out with the overflow of the information. In the context of this thesis, visualization refers to provide a visual representation that allows a human to observe visual traits of interest or relevance within a set of image data Keller *et al.* (1994), i.e. image regions related to labels of interest such as age or disease.

Visualization of medical image data McAuliffe *et al.* (2001) can be generally based on an individual image or group of image data, e.g. a single MRI brain scan or a collection of such scans. Individual image visualization is useful in the case where we seek to interpret the scan of an individual patient, e.g. personalized medicine. Group-wise visualization is useful in the study of a population, in order to understand the link between image structure and labels across a group of patients.

In the case of individual images, we seek visualization that can provide complementary information above and beyond simple classification. By providing a visualization of the spatial layout of features used in classification, it may be easier for a clinician to understand or quantify the uncertainty of classification score.

In order to address the need for human interpretation, we consider scan interpretation not as an automatic classification task, but rather as a visualization task, where the goal is to generate an informative visual summary of pertinent disease-related image information. Rather than providing fully automatic classification, we seek to produce a concise visual summary of disease-related information that allows a human to make rapid, yet well-informed decisions, Figure 1.10. Data visualization is considered as an important communication tool between machine and human works.

Zhang *et al.* (2017) provide an attention-based visualization mechanism that highlights relevant image structure. In contrast, our visualization highlights not only relevant structure, but also class conditional highlighting with color channels.

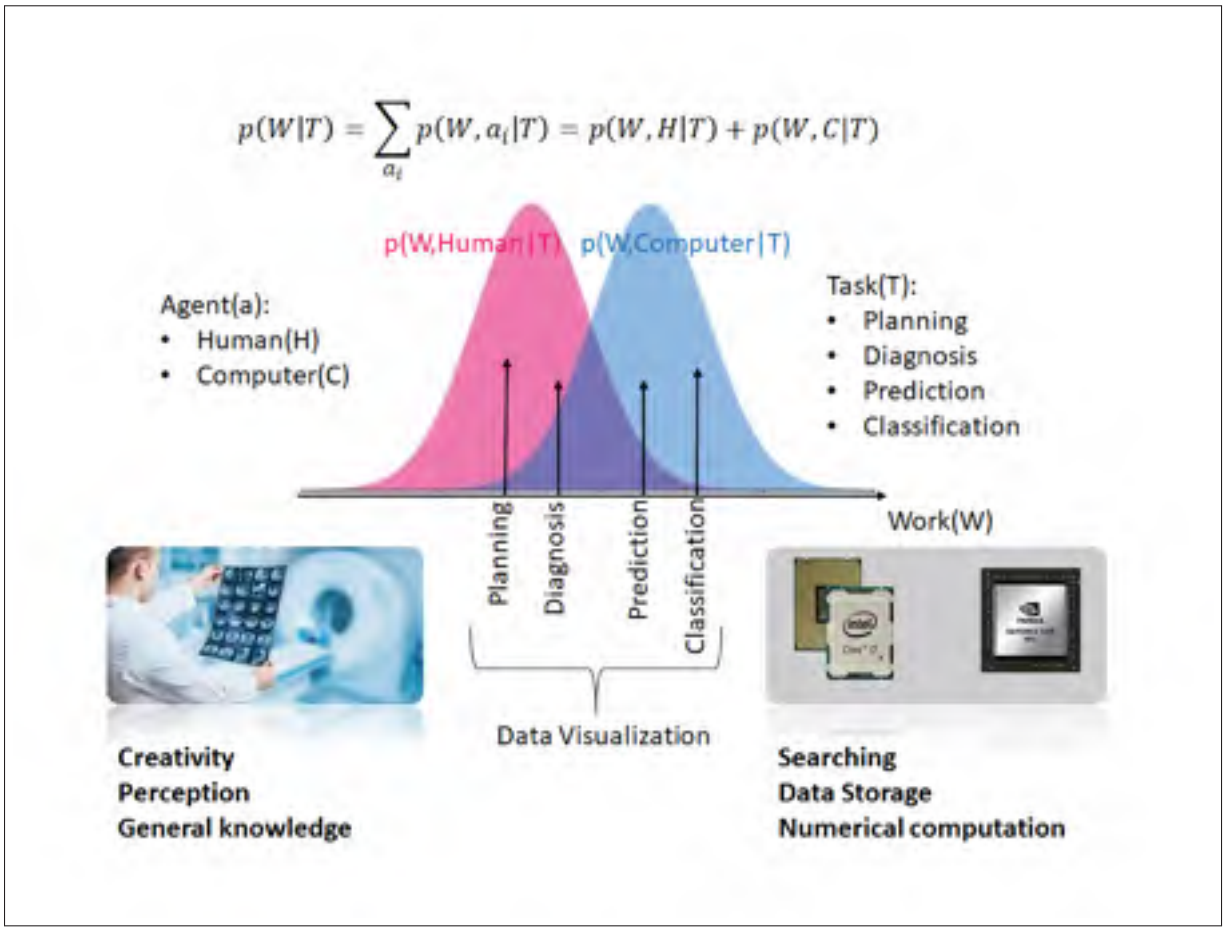

Figure 1.10 An illustration of importance of data visualization, which is considered as an important communication tool between machine and human experts these days. Data visualization could affect and increase the overlap of two joint probability distributions given by specific tasks, $p(W, H|T)$ and $p(W, C|T)$

In many fields of medical and clinical research, imaging has become an essential ingredient. Recently, computers equipped with the graphics software and hardware, make it possible to visualize the internal organs easily. This work introduces a visualization program specifically designed to meet the needs of patient's age identification from their brain's MRI in medical research community.

People can understand the significance of data through data visualization which places data in a visual context. For example, one of the confusion that might be occurred by text-based data is that many patterns and correlations might go undiscovered. By data visualization software,

patterns and trends can be strongly recognized easier. Various algorithms attempt provide a visual summary of image collections, Simon *et al.* (2007), including localized information regarding visual traits, e.g. image patches indicative of face gender, Toews & Arbel (2009). There has been little work in providing a visualization for the purpose of decision making.

1.3.1 Group-wise Visualization

Group-wise visualization aims to highlight the spatial layout of features most related to the label of interest, e.g. age or disease, across a population. Group-wise data visualization let us to see trends, patterns, changes, differences and similarities clearly between different groups. In medical imaging, this topic would be important as it could help medical scientist to discover the disease's features by seeing to the images.

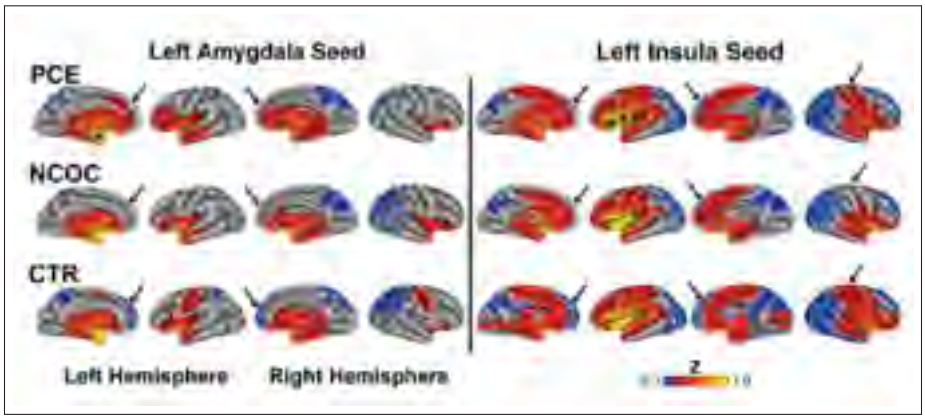

Figure 1.11 An example of group-wise data visualization,
Similarities and Differences visualization of functional connectivity
in brain regions across a specific group
Taken from Salzwedel *et al.* (2015)

For seeing the class-related features voxel-wise visualization methods are widely used, e.g. voxel-based morphometry (VBM). In brain's MRI studies, most of the examinations use voxel-based morphometry method which characterizes group's brain regional volume and different concentration of brain's tissue. Ashburner & Friston (2000) proposed an approach to estimate the local amount of specific tissue based on group examination and generate voxel-wise visualization.

This voxel-wise estimation method can not be used for individual image visualization or classification. VBM can be used for examining the gray matter and also white matter. VBM conceptually contains the following processing steps which might obtained by the statistical analyzing:

• Tissue Classification,

• Spatial Normalization,

• Spatial Smoothing.

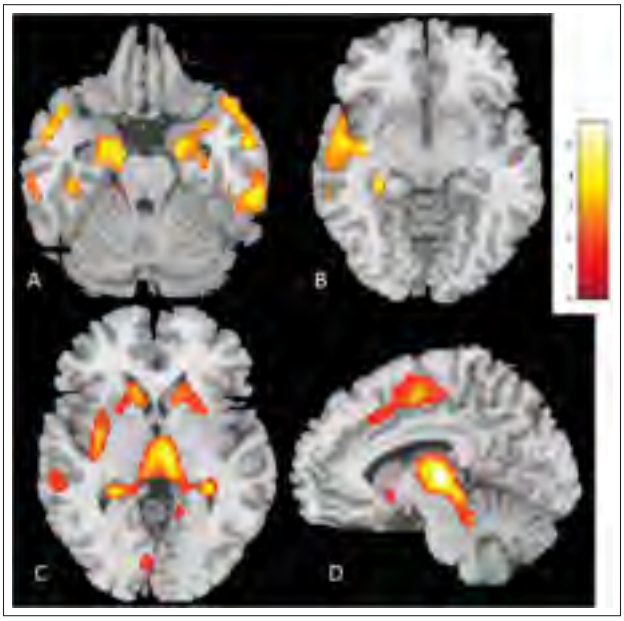

Figure 1.12 An example of Voxel Based Morphometry visualization survey in comparison of epilepsy patients and aged-match controls
Taken from Chang *et al.* (2012)

Tissue classification was presented by Ashburner & Friston (1997) previously. The non-uniformity of image intensity which have been arises in MR imaging, were corrected by this method. As it is mentioned above, all images are spatially normalized to the same stereo space in VBM. In other word, all images aligned to the same template-model image, by reducing the sum of squared differences between them. The spatial normalization does not achieve an exact match between every cortical feature, but entirely works in global brain shape differences. After

smoothing the extracted gray matter and also tissue localization by statistical analysis group differences would be discovered. This method's return is a statistical parametric map which refers to the significant differences of gray matter regions between different groups.

1.3.2 Image-wise Visualization

Single-subject visualization is similar to group data-driven visualization. But there are important differences between these approaches too.

Toews *et al.* (2010) presented a method named Feature-Based Morphometry, FBM, to find the unique anatomical patterns in 3D images in a fully data-driven manner. For analyzing the features in characteristic scale, the scale-space theory has been used. The features are generalized by a probabilistic model in terms of their geometry, appearance, and relation to subject group. Learned features are related to group-related anatomical structure based on the probability of features happening within a specific group. FBM avoid one-to-one inter-subject by modeling the image as a collage of local invariant features that do not need to present in all subjects.

FBM identifies features invariant to changes in image scales. It uses scale–invariant feature transform (SIFT) technique, Lowe (2004), in 2D photographic imagery. Features are extracted in a Gaussian scale-space since feature geometries are independent to resolution and reflect location and scale of the image. As it is hard to identify features which are come from the same underlying tissue in different subjects, FBM uses probabilistic model to identify features in terms of their appearance, geometry, and relationship to subject group. FBM tries to avoid one-to-one correspondence and discovers pattern of anatomical structure as distinctive scale-invariant feature, Lowe (2004), Mikolajczyk & Schmid (2004) based on the statistical probability correspond to subsets of subjects.

The first step of FBM is to extract features and indicate the location and size of salient anatomical patterns by localizing them in image scale and space. One approach to identify image patterns with characteristic of scale and size independent to image resolution, is to detect features that are invariant to scale, Lindeberg (1998). Gaussian scale-space is the most common kernel for

this application.

$$I(x, \sigma) = I(x, \sigma_0) * G(x, \sigma - \sigma_0) \tag{1.5}$$

Equation 1.5 represents the convolution of image I and Gaussian kernel at location x and difference of the scales σ and σ_0.

The SIFT algorithm was generalized to 3D volumetric data video processing Scovanner *et al.* (2007), 3D baggage scanners Flitton *et al.* (2010), and medical image data, Cheung & Hamarneh (2009); Allaire *et al.* (2008); Toews & Wells (2013). The work in this thesis is based on the 3D SIFT-Rank Toews & Wells (2013). In this work, full 3D orientation (i.e. a 3-parameter 3x3 rotation matrix) is estimated from unbiased spherical histograms, rather than solid angle angles Scovanner *et al.* (2007); Allaire *et al.* (2008) or incomplete 2D information Cheung & Hamarneh (2009).

This method has been applied in a wide variety of tasks, including modeling infant brain development in Toews *et al.* (2012), inter-modality keypoint matching Toews *et al.* (2013), lung CT scan alignment Gill *et al.* (2014); Toews *et al.* (2015) of 20000 subjects, whole-body medical image segmentation via keypoint transfer segmentation Wachinger *et al.* (2015, 2018), 4D cardiac ultrasound matching Bersvendsen *et al.* (2016) identifying brain MRIs of family members Toews & Wells (2016), and large-scale multi-modal analsys of T1, T2, diffusion and functional MRI modalities Kumar *et al.* (2018)). Alternative keypoint descriptors have been proposed for diffusion MRI Chauvin *et al.* (2018). The Jaccard distance between bag-of-feature sets can be used to identify family members including twins and siblings, and detected errors in large MRI datasets Chauvin *et al.* (2019, 2020). The 3D SIFT-Rank method was also used for robust 3D ultrasound image alignment in the context of image-guided neurosurgery Luo *et al.* (2018c,a), including non-rigid registration Machado *et al.* (2018b); Frisken *et al.* (2019a,b).

The second step is to define a probabilistic model to associate the features of different subjects to subject groups. Due to the difficulty in finding features of same underlying tissue in different subjects and inter-subject and inter-group variability, the probabilistic model has been used. At first, the features align approximately to an atlas or a reference by a geometrical transform T

which is constant in probability model. Each feature denoted as $f = \{a, \alpha, g, \gamma\}$. In which the a is the vector of image measurement which shows the appearance of local feature. a shows that if the a is a valid feature instance or a valid noise. $g = \{x, \sigma\}$ shows the geometry of feature, in which the x and σ represent the location and scale respectively. γ indicates the presence or absence of geometry in subject image. If C represents the class from which the subjects are sampled and F represents a set of local feature $F = \{f_1, ..., f_N\}$ the probability of $p(C, T|F)$ is:

$$p(C, T|F) = \frac{p(C, T)p(F|C, T)}{p(F)} = \frac{p(C, T)\Pi_i^N p(f_i|C, T)}{p(F)} \tag{1.6}$$

In this equation $p(F)$ is the probability of feature set F. $p(C, T)$ is the joint probability of class and geometrical transform, and $p(f_i|C, T)$is the probability of feature f_i.

Learning procedure consist of clustering features in terms of their appearance(a) and their geometry(g). Features related to each cluster show the different observation same anatomical structure. g and a are two different cluster. Features belonging to these clusters are similar to each other in terms of geometry and appearance respectively. f_j is similar to feature f_i in terms of geometry if their location and scale differ by less than the threshold ϵ_x(geometry location distance threshold) and ϵ_i (geometry scale distance threshold). The scale difference is in log domain and the location difference is normalized by the feature scale σ_i.

Choosing values larger than these errors causes different anatomical structure to be grouped together. The maximum acceptable deviation for appearance similarity is ϵ_a(appearance distance threshold) . For each feature f_i , the function of ϵ_{ai} has been defined by the Euclidian distance between appearance vector a_i and a_j less than the threshold ϵ_{ai}: The new feature f_j in image is detected if its location and scale differ less than error with the model feature f_i . The classification expressed by the following equation where C and $\bar{C}$ shows different class labels:

$$C^* = argmax_C\{\frac{p(C, T|F)}{p(\bar{C}, T|F)}\} = argmax_C\{\frac{p(C, T)}{p(\bar{C}, T)}\Pi_i \frac{p(f_i|C, T)}{p(f_i|\bar{C}, T)}\} \tag{1.7}$$

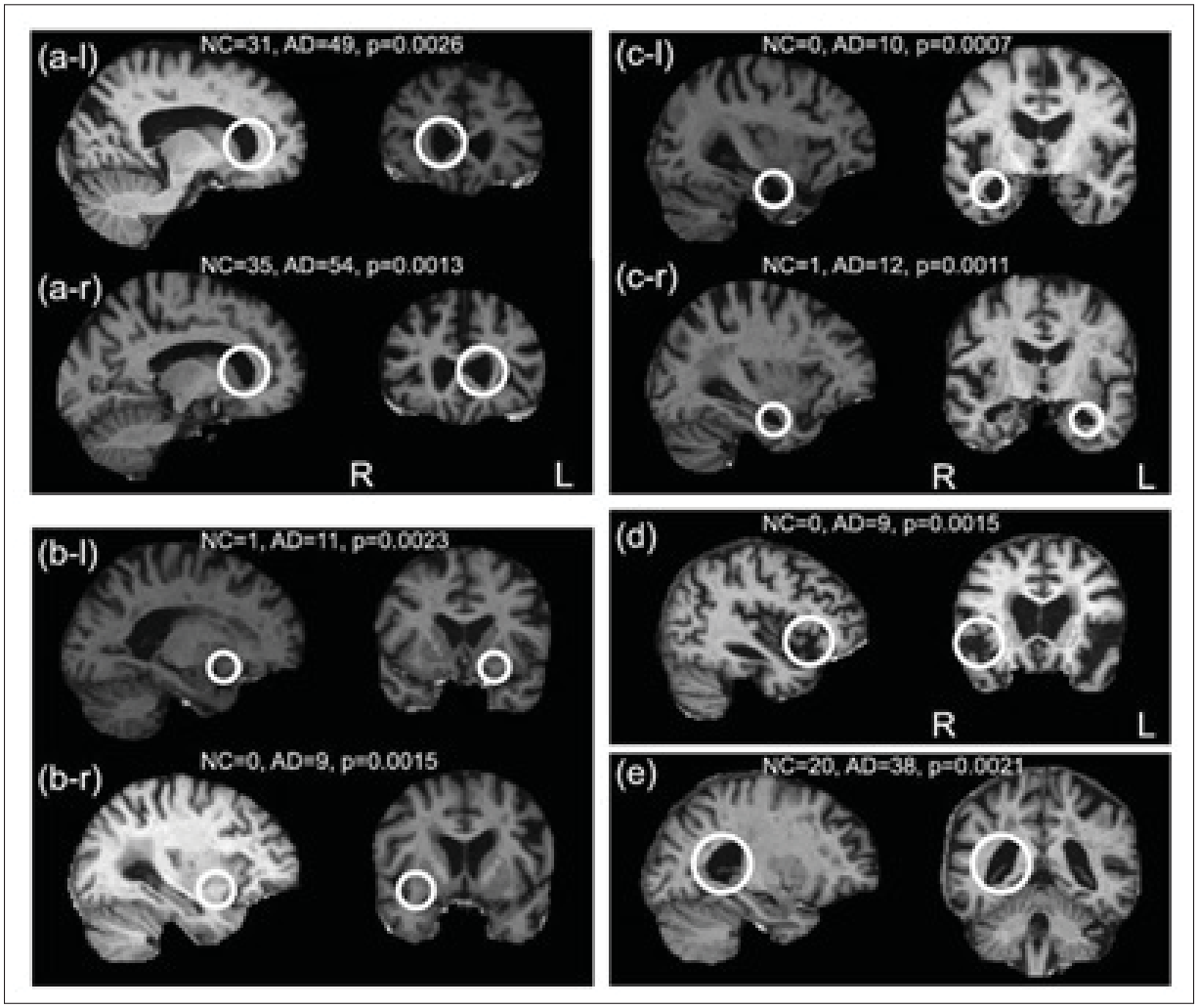

Figure 1.13 Representing similar anatomical regions by visualizing the most significant AD-related features."Examples of the eight most significant AD-related features, shown in sagittal and coronal slices. The feature occurrence frequencies within 75 AD and 75 NC subjects and associated uncorrected p-values are given. Out of eight features, six represent similar anatomical regions identified independently in left and right hemispheres. All represent neuroanatomical regions known to be affected by AD"
Taken from Toews *et al.* (2010)

1.4 Hypothesis Testing

1.4.1 p-value

During recent years, data collection from various research area like genetic, biochemistry and imaging of the brain increased dramatically. Statistical test analysis needs to be performed on the various test with different variables. Therefore, to address this issue, a new concept has emerged called multiple testing Bender & Lange (2001), Olejnik *et al.* (1997). Multiple tests investigates simultaneously several hypothesis. It is performed while we are dealing with a large number of data, thus it is very important to improve the methods in order to increase the accuracy and validity of the results obtained by research and tests. The assumptions and the level of statistical significance in the test is usually based on the calculations related to P values and error type 1.

In this test calculating p-value can address the validity of zero assumption. For this purpose, the first type error (a) is considered as a maximum threshold of p-value. If the observed value is less than the amount set with probability (1-a), the assumption of zero will be rejected, while the definition of statistical significance level test in the multiple test is a complex issue. The amount of the erroneous findings is significantly increased by the improper handling of individual analysis and ignoring the relationship between assumptions Meng *et al.* (1994).

A set of data samples of two or more parameters might be available in many contexts, e.g. image features and labels such as age or disease. Hypothesis testing is used to quantify the significance of potential correlations between these parameters, e.g. co-occurring image pattern and labels.

Hypothesis testing involves evaluating the probability $p(D|H_0)$ of an observed data on a set of D, under a null hypothesis of H_0 or model assuming uncorrelated parameters Fisher (1992). This probability is known as a p-value. In case of correlation, parameters tends to be low, since the data follows an alternative hypothesis model. In the case that the p-value is below a certain threshold or significance level, commonly 0.05 or 0.01, we can reject the null hypothesis.

For example, consider the case where the data consist of joint observations of an image feature $\{F, \bar{F}\}$ and a binary image class label $\{C, \bar{C}\}$. Such a dataset can be summarized by a 2x2 contingency table describing the joint probability of feature presence or absence vs. class label, e.g. $p(F, C)$.

Fisher's exact test Fisher (1992) evaluates the probability of observing the table under a hypergeometric distribution, i.e. the null hypothesis that feature presence is independent of the class label, and a uniform joint distribution $p(F, C|H_0)$. The p-value is thus high in the case of uncorrelated feature and class labels, when a feature and class labels are highly correlated, the p-value or data probability is low.

In the case where multiple data samples are available, e.g. many features, it is important to correct significance or p-values for multiple comparisons, as a certain percentage of low p-values are expected due to random chance. Correction can be performed by adjusting the p-value threshold, Abdi (2007). An alternative to correcting p-values for multiple comparisons is the false discovery rate(FDR),Benjamini & Hochberg (1995).

The FDR has been previously used to quantify the group-wise Type I error rate,Toews *et al.* (2010). The essence of our proposed method is to use significance testing and the FDR as a parameter estimation technique.

A theoretical statement concerning a certain feature of the studied statistical population is called hypothesis. A procedure of assessing whether sample data is consistent with statements (hypotheses) made about the statistical population is defined as a hypothesis testing or significance testing Banerjee *et al.* (2009). In a brief explanation, a decision about the hypothesis is made based on sample data. We seek to answer questions like "Is there a difference between the samples or Is there a relationship between the variables".

When the true effect of each test is nonappearance, the error of result can be increased. By increasing the number of test, this error and false probability is also increased. Controlling the

system error can be achieved by false discovery rate which leads to achieve significant p-value. Benjamini & Hochberg (1995) introduced the false discovery rate as a procedure to control false.

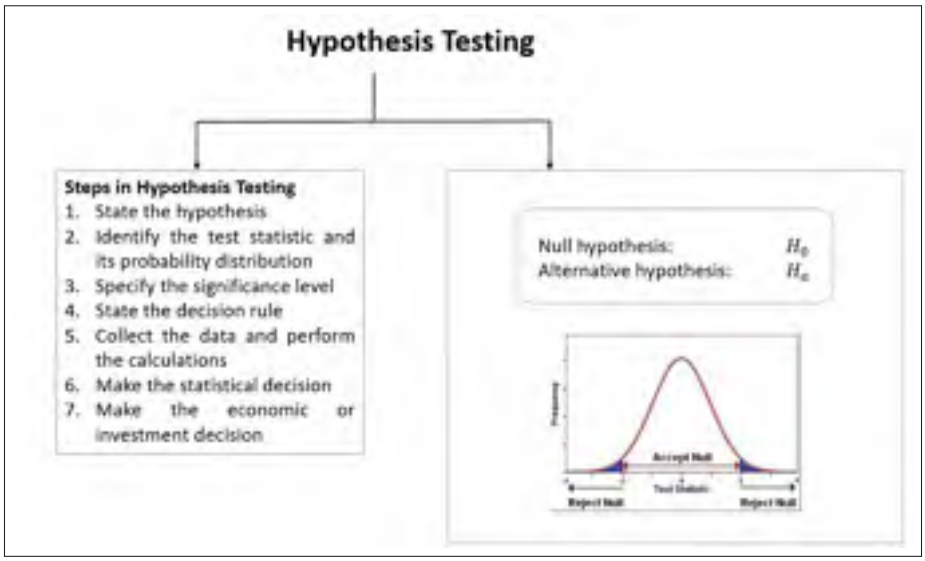

Figure 1.14 Steps in hypothesis testing

Neyman & Pearson (1933), Shen *et al.* (2018), published a paper that described an approach to decide between competing hypotheses without requiring prior beliefs. This was the most frequent basis for hypothesis testing. They were not particularly concerned with the confidence that one should place in any specific scientific hypothesis, but presented a rule-based framework, describing Type $I(\alpha)$ errors (falsely rejecting the null hypothesis) and Type $II(\beta)$ errors (falsely rejecting the alternative hypothesis). Whether each hypothesis is true or false, there should be a rule to behave them Biau *et al.* (2010).

They discarded the p-value, and became bitter rivals with Fisher. While Fisher feuded with Neyman and Pearson over the next two decades, an amalgam of the methods began to appear in statistics textbooks. In time, the p-value came to be misinterpreted as a dichotomous marker of statistical significance, Biau *et al.* (2010).

CHAPTER 2

METHODOLOGY

This thesis seeks a method for visualizing the posterior probability of a class label across the image space, for example the probability of disease across the MRI of a patient acquired at a hospital, thereby offering additional information regarding the spatial layout of the class across the image for visual inspection by a health professional. It seeks a learning-based method, that is capable of leveraging an arbitrarily large set of image data, e.g. brain MRI, to continuously improve as data arrive.

In this chapter, we present our method, including the local features used in analysis in section 2.1, the generative model linking local features to class in section 2.2 and the distance metrics and estimation of model parameters in section 2.3 and section 2.5 respectively. And also, visualization technique is illustrated in 2.6 and classification computation is expressed by 2.4.

2.1 Local Feature Format

Standard 3D medical image data sampled on a 3D coordinate lattice $\bar{x} \in \mathbb{R}^3$ require a large amount of memory to represent. Our work considers these data in the local salient feature format, where data are sampled at sparse keypoints the 7 dimentional manifold of 3D location, 3D orientation and 1D scale $\{\bar{x}, \Theta, \sigma \in \mathbb{R}^7$. All work on this thesis is based on the implementation of 3D SIFT keypoint format Toews & Wells (2013)[1] Toews & Wells (2009) derived from the widely used scale-invariant feature transform, Lowe (2004). Feature detection methods are under continuous development, and recent methods use GPU-based convolution operatorsOno *et al.* (2018). The SIFT method is a classic method based on a highly efficient recursive convolutional neural network structure, designed for use on a CPU rather than a GPU, using a single Gaussian filter per layer. A major advantage of the SIFT approach is that the resulting feature set extracted in scale-space is mathematically invariant to global variations in image geometry and appearance

[1] www.matthewtoews.com

e.g. due to image misalignment or intensity changes. Other advantages is that the filtering process is based on filters of a specific computational form, and thus unbiased by the particular training dataset. These informative image patterns are unbiased with the individual subject images. In this study, the maximum of salient benchmarks computed through the images could be considered as interesting local features. Despite of existence of a variety methods of features detection Bay *et al.* (2006), we extract them by a 3D Generalization of the SIFT algorithm, Toews & Wells (2013) insomuch as SIFT algorithm has many distinctive advantages such as being robust to occlusion and clutter, and also having a very good recall rates. In this thesis image local features are considered as latent random variables.

An essential part of medical image processing is image registration,Maintz & Viergever (1998). Using image registration, the patient's data have been spatially normalized into a standard reference space. In this work, the images are aligned by using global linear registration method which generate approximate inter-subject registration. Affine registration to Talairach stereotaxic space is considered as a famous type of global linear transformation for brains, Talairach & Tournoux (1988). Although the goal of affine registration is not to have one-to-one inter-subject matching, there is alignment between corresponding anatomical structures in different subjects. Since the purpose of this work is to extract the important group related differences, it is important to focus on local image patterns and avoid of any over-fitting. Toews & Wells (2013) proposed a method to focus on the maximum a-posterior (MAP) transform $T_{MAP} = \underset{T}{\text{argmax}} p(T|I)$, maximizing the posterior probability of T conditional on I. Where the global linear transform mapping of image I to a standard reference space is represented by T.

Feature extraction is based on the Gaussian scale-space representationLindeberg (2013). Features are localized in space and scale which are related to the location and size of the distinctive anatomical patterns respectively. The location and scale of the distinctive features of registered images, $(\bar{x_i}, \sigma_i)$, are acquired by calculation of the extreme of a difference-of-Gaussian (DoG), Equation 2.1. Detected features are considered as an spherical image regions centered on location $\bar{x_i}$ and a radius which is proportional of the scale σ_i. In Equation 2.1, $f(\bar{x_i}, \sigma)$ is the convolution of the image I with a Gaussian kernel of variance σ^2 and k is the multiplicative

sampling rate, Lowe (2004).

$$\{(\bar{x}_i, \sigma_i)\} = local \underset{\bar{x}_i, \sigma}{\operatorname{argmax}} |f(\bar{x}_i, k\sigma) - f(\bar{x}_i, \sigma)| \tag{2.1}$$

Detected salient points are reoriented and re-scaled to the fixed size (here 11^3 voxels) and transformed into a gradient orientation histogram (GoH) representation over 8 spatial bins and 8 orientation bins. It results appearance feature descriptors in 64-elements. Ultimately, by rank-ordering, Toews & Wells (2009), descriptors are transformed into an ordinal representation, providing resistance to monotonic changes. Consequently, the elements take their rank in an array which are sorted according to GoH value.

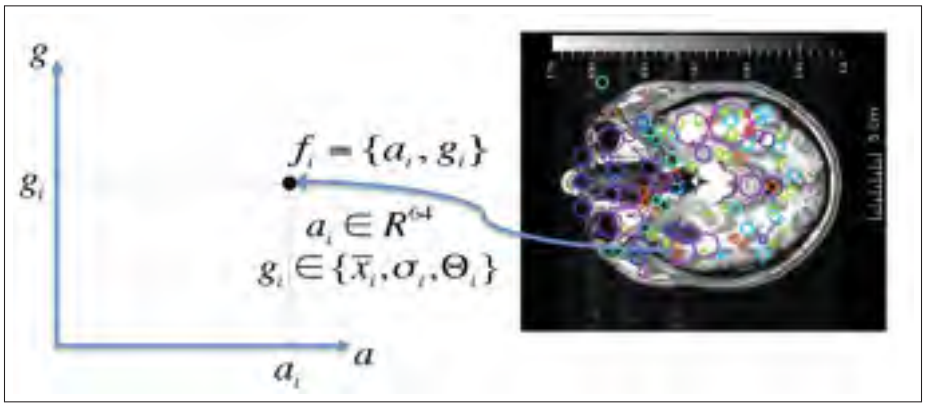

Figure 2.1 Illustrating the space of local feature data $f = \{a, g\}$ consisting of joint appearance a and geometry g subspaces. Each feature sample $f_i = \{a_i, g_i\}$ consists of a geometrical component g_i and an appearance a_i component, here illustrated as two scalar dimensions for visualization purposes. The geometry $g_i = \{\bar{x}_i, \sigma_i, \Theta_i\}$ consists of a 3D location $\bar{x}_i$, a scale σ_i and an orientation matrix Θ_i. The appearance a_i is a descriptor, here a 64 dimensional SIFT-Rank vector

Figure 2.1 illustrates the space of local features. Each feature consists of geometrical and appearance components. The size of circles are determined by the scale of each feature. Dataset using by this research has many such a salient key point as the observed features.

2.2 Generative Model

The main goal of this method is to determine kernel density parameters such that the statistical significance of kernels is maximized.

Classification approach in this work seeks to provide an estimate of the posterior probability of class C conditional on feature F and spatial location $\bar{x}$. Let C be a discrete random variable of class, e.g. $C = \{Old, Young\}$, $\bar{x} = \{x, y, z\}$ be a random variable of 3D location and $F = \{f_1, \ldots, f_i, \ldots, f_N\}$ be a latent local feature variable defined over N features extracted in an query image. Using Bayes rule of conditional probability, the posterior probability may be expressed as

$$p(C|F, \bar{x}) = \frac{p(F|C, \bar{x})p(C|\bar{x})}{p(F|\bar{x})} \propto p(F|C, \bar{x})p(C|\bar{x}), \tag{2.2}$$

where $p(C|F, \bar{x})$ represents the posterior probability of class C conditional on observed feature F and spatial location $\bar{x}$. The main approach of this work is to visualize this posterior probability of the class $p(C|F, \bar{x})$. In our work, F is assumed a set of independent, identically distributed (IID) random variables of observed features f_i, $p(F|C, \bar{x})$ can be defined as

$$p(F|C, \bar{x}) = \prod_{i}^{N} p(f_i|C, \bar{x}) \tag{2.3}$$

While in general, features in F may exhibit a degree of covariance, e.g. between spatial neighbors within an image, the assumption of independence in Equation 2.3 results in a computationally efficient model, and is reasonable due to the fact that they represent unique image structure localized in scale and space, and no direct functional relationship. In Equation 2.3, $p(f_i|C, \bar{x})$ represents the posterior probability of variable feature f_i given class C and spatial location $\bar{x}$. As data consists of discrete feature samples as illustrated in Figure 2.2, $p(f_i|C, \bar{x})$ may be represented using a kernel density estimator as the marginalization of latent random feature variable $\{f_j\}$:

$$p(f_i|C, \bar{x}) = \sum_{f_j} p(f_i, f_j|C, \bar{x}) \approx \sum_{f_j \in kNN(f_i)} p(f_i, f_j|C, \bar{x}) \tag{2.4}$$

In Equation (2.4), the sum over f_j is computed across all previously observed features stored in memory. In practice, kernel values $p(f_i, f_j|C, \bar{x})$ computed based on observed feature f_i are low (approximately zero) with the exception of a set of nearest neighbors $kNN(f_i) = \{f_j\}$, the set of the k-nearest neighbors of f_i in joint feature space $f = \{a, g\}$, due to the sparsity of SIFT features. The sum in Equation (2.4) may thus be approximated for efficiency by a sum over a small set of $kNN(f_i) = \{f_j\}$ features for which $p(f_i, f_j|C, \bar{x})$ is significantly non-zero.

We seek to estimate the parameter ϵ_i which is defined as size of the band-width shown by Figure 2.2 contains informative nearest neighbors.

$$p(f_i, f_j|C, \bar{x}) = \frac{p(C|f_i, f_j, \bar{x})p(f_j|f_i, \bar{x})p(f_i|\bar{x})}{p(C|\bar{x})}, \tag{2.5}$$

$$= \frac{p(C|f_j)p(f_j|f_i)p(f_i|\bar{x})}{p(C|\bar{x})}. \tag{2.6}$$

Equation 2.5 is due to the chain rule of probability. Under the assumption that $p(C)$ is constant, e.g. all variable of class are equally probable a priori and also class C and spatial location $\bar{x}$ are conditionally independent given feature f_i. Intuitively, given an invariant feature f_i, spatial image location offers no additional information regarding C.

Assuming $p(f_i)$ and $p(\bar{x})$ are constant e.g. all spatial locations and variable of observed features are equally probable a priori, conditional density $p(f_i|\bar{x})$ can be expressed as

$$p(f_i|\bar{x}) = \frac{p(\bar{x}|f_i)p(f_i)}{p(\bar{x})} \propto p(\bar{x}|f_i) \tag{2.7}$$

Substituting Equations 2.7 and 2.6 into Equation 2.2 leads to this problematic model:

$$p(C|F, \bar{x}) \propto \prod_{i}^{N} \sum_{f_j \in kNN(f_i)} p(C|f_j)p(f_j|f_i)p(\bar{x}|f_i) \tag{2.8}$$

$p(C|f_j)$ is a kernel density function model across a large set of training data, i.e. feature sets F and corresponding variable labels C.

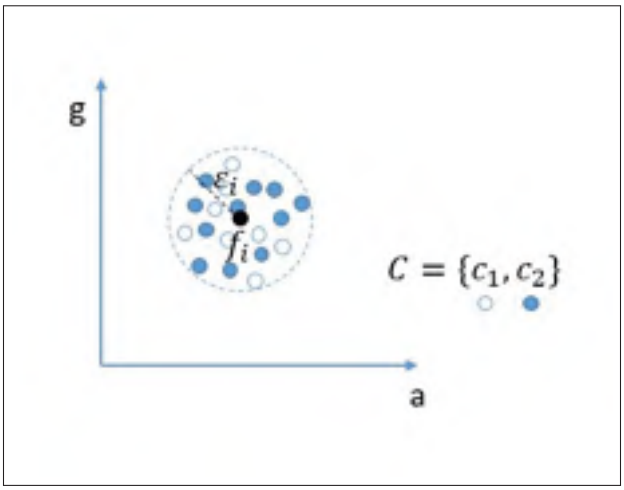

Figure 2.2 Illustrating classification via density estimation in the space of local features

The probability density $p(f_j|f_i)$ can be factored into conditional densities over appearance a and geometry g as an image feature $f = \{a, g\}$ is represented in statistically independent terms of an appearance descriptor a sampled according to the geometry g. In a typical image, g is a pixel coordinate and a is an image appearance intensity measurement. Therefore, $p(f_j|f_i)$ may be expressed as Equation 2.9:

$$p(f_j|f_i) = p(a_j|f_i)p(g_j|f_i) = p(a_j|a_i)p(g_j|g_i) \tag{2.9}$$

Here we consider the scale-invariant keypoint representation, i.e. the 3D scale-invariant feature transform format, Toews *et al.* (2010). In this format, each key-point or feature $f_i = \{g_i, a_i\}$ is defined as a geometry descriptor $g_i = \{\bar{x}_i, \sigma_i, \Theta_i\}$, where $\bar{x}_i$ is 3D location, σ_i is scale and Θ_i is orientation, and an appearance descriptor a_i computed from the image content about the key-point location. σ_i shows scale of the feature. By ignoring the orientation, we may have:

$$p(g_j|g_i) = p(\bar{x}_j|\bar{x}_i)p(\sigma_j|\sigma_i) \tag{2.10}$$

Posterior probability of class given by feature f_j indicate the class estimation which is based on the nearest neighbors corresponding class. Also, $p(f_j|f_i)$ points to the nearest neighbors algorithm. Where n is the number of informative features per subject/image and k is the number of informative nearest neighbors of observed informative feature of the subject. Finding these two parameters are the main objective of this work. Let's take a look at them in the following parts.

Posterior probability of spatial location given by observed feature, $p(\bar{x}|f_i) = G(\bar{x_i}, \sigma_i)$ is a Gaussian density with parameters mean $\bar{x_i}$ standard deviation σ_i quantifying the spatial extent of feature f_i. Term w_i refers to the weight which defines the intensity of color and brightness of the point in the visualization technique.

$$p(\bar{x}|f_i) = \sum_{i}^{n} w_i e^{-\frac{(\bar{x} - \bar{x_i})^2}{2\sigma_i^2}} \tag{2.11}$$

2.3 Distance Metric

The k-nearest neighbors classification method is used in this work. Before any data-based decision, it is important to determine an appropriate distance metric measurement among variables and their neighbors to recognize better the input data pattern. Data used by this experiment contains high dimensional features descriptors and their corresponding labels. These features descriptors are sorted in two categories: appearance and geometry. Here, we are going to visualize them in the feature-space and calculate their distances to find significant nearest neighbors.

In general, the Mahalanobis distance may be used to estimate the distance from a descriptor sample f_i to the mean μ_i of a multi-variate Gaussian density $G(\mu_i, \Sigma_i)$ of covariance Σ_i, defining hyper-eliptical isodistance contour in the descriptor space Mahalanobis (1936) . Under the assumption of independent and identically distributed descriptor elements, the covariance matrix Σ_i is diagonal and the Mahalanobis distance becomes equivalent to the Euclidean distance, one

of the most commonly used distance metrics. The Euclidean distance is also know as the $L2$ norm or Minkowski metric with $p = 2$.

The model in this thesis considers a distance metric combining feature appearance and geometry components as orthogonal dimensions

$$d^2(f_i, f_j) = d_a^2(f_i, f_j) + d_g^2(f_i, f_j) \tag{2.12}$$

where d is the distance between observed feature f_i and its neighbors $f_j \in kNN$. Note that orthogonality of Gaussian random variables also implies independence. Since each keypoint includes appearance and geometry descriptors we may have:

$$d_a(f_i, f_j) = d(a_i, a_j) = ||a_i - a_j|| \tag{2.13}$$

$$d_g(f_i, f_j) = d(g_i, g_j) = ||g_i - g_j|| \tag{2.14}$$

Therefor d_a and d_g in Equations 2.13 and 2.14 refer to the notations of euclidean distance metric between the observed feature's appearance and geometry descriptors and their neighbors.

In order to compute the geometrical distance in this experiment, only the location parameters (x, y, z) are considered. Each feature has their own size or scale which makes uncertainty to localize them accurately. To see the effect of feature's scale there would be a new definition of geometry distance metric which includes the euclidean distance of descriptors in normalized image space as well. In other word the new defined variance is proportional to the product of feature scales, $\sigma_i^2\sigma_j^2$.

$$d_g(f_i, f_j) = d(g_j, g_i) = \frac{||g_j - g_i||}{\sigma_j^2\sigma_i^2} \tag{2.15}$$

Where σ_j relates to the observed feature's neighbor's scale. Finally, geometry distance metric is scale normalized of euclidean distance of feature's location descriptors, Eq. 2.15.

Figure 2.3 illustrates feature-distance space (FDS), with observed feature located at the origin, the point $(0, 0)$. The other features shown in this graph are sorted based on their appearance and geometry descriptors distances with observed feature, d_a and d_g. In order to compare Geometry and appearance distance with each other they need to have unit Gaussian. In order to make them normalized, distance measures have to be divided by their standard deviation. σ_a and σ_g are respectively d_a's and d_g's standard deviation.

$$\hat{d}_a = \frac{d_a}{\sigma_a}, \hat{d}_g = \frac{d_g}{\sigma_g} \tag{2.16}$$

After normalization, it is needed to find one parameter distance metric between f_i and f_j. Therefore, the following formula is as follow:

$$d = |f_i - f_j| = \sqrt{\hat{d}_a{}^2 + \hat{d}_g{}^2} \tag{2.17}$$

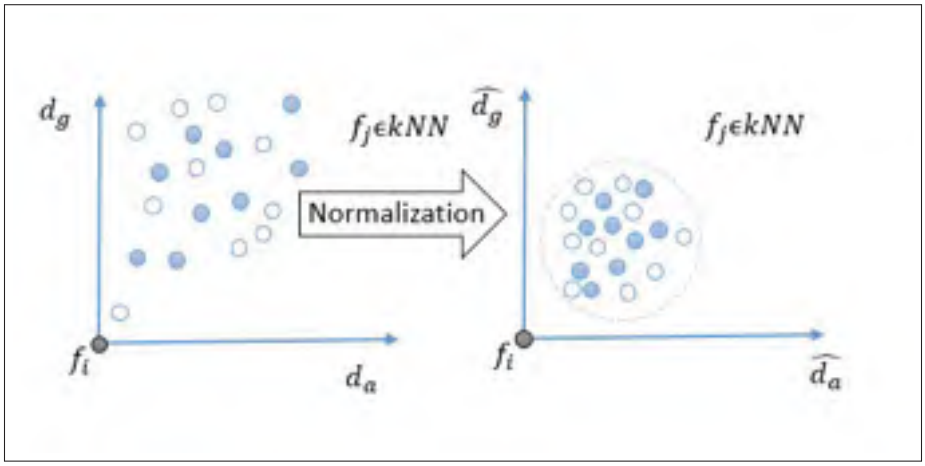

Figure 2.3 Normalization of features in Feature-Distance (FD) space

Figure 2.4 illustrates the final distance metric which is shown as a radius of each orange dashed quadrant. Any features on the smaller quadrant is closer to the observed feature. This is shown by d in equation 2.17. The nearest neighbors are achieved by sorting this distance measures in ascending order .

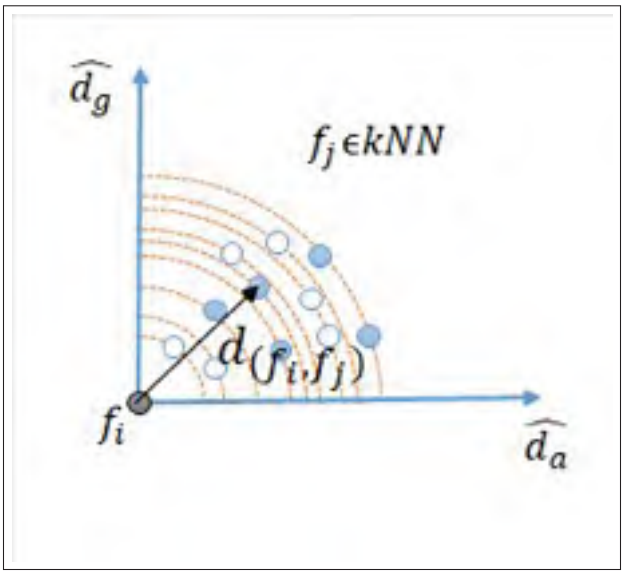

Figure 2.4 Feature-Distance space; Vector $d_{(f_i,f_j)}$ shows the distance between observed feature f_i and its neighbor f_j

2.4 Classification

Classification considers a model of conditional feature independence $p(F|C) = \prod_i p(f_i|C)$, where individual feature densities $p(f_i|C)$ may be generally modeled using kernel density estimation (KDE). For the purpose of classification, omit spatial variable $\bar{x}$.

The theoretical contribution of this work is a kernel density estimator based hypothesis testing, where $p(f_i|C)$, the kernel bandwidth is set such that the probability of observing a set of neighbors given independence $f_i \perp\!\!\!\perp C$ is minimized. Therefore for a feature f_i, we identify a set of k nearest neighbor features $N_i = \{f_j : f_j \in kNN(a_i) \cap f_j \in kNN(g_i)\}$ in both appearance and geometry, where f_j is a training feature with associated class label C_j.

The probability of randomly selecting the feature set N_i assuming a uniform or uninformative distribution over class C is modeled by the hyper geometric distribution. The probability of observing a set where the association between features and class C is as strong or stronger than the discrete distribution defined by N_i may be evaluated using Fisher's exact test.

The most informative features are those one that have the lowest $p(f_i|C,\bar{x})$. This low probability density rejects the null hypothesis and proves non uniformity distribution over the class. To calculate fisher's exact test, a discrete probability distribution such as hyper geometric distribution could be used to describe k relative features among n relative and non-relative features (k successes in n draws).

Algorithm 2.1 describes the steps in computing minimum p-value bandwidth thresholds for a new set of feature data $\{f_i\}$, given a discrete nearest neighbors $\{f_j\}$ and binary labels C Here we suggest training classifier via a supervised learning technique. Function of p-value is calculated based on contingency table which is explained with more details in the following section. It is assumed that the classification type for this pseudo code is binary classification.

Algorithm 2.1 Estimating minimum p-value thresholds $\{minimum\_p\_value_i\}$ corresponding to an input set image features $\{f_i\}$ and a set of training features $\{f_j\}$ and labels C in memory.

Input Input feature set $\{f_i\}$, training features $\{f_j\}$
Output Set of min p-values $\{minimum\_p\_value_i\}$
foreach f_i **do**
    find k nearest neighbors; $f_j \in kNN(f_i)$, $j \in \{1, ..., k\}$
    $p_{min} \leftarrow 1$
    forall $f_j \in kNN(f_i)$ **do**
        check if $p\_value(f_i, f_j) < p_{i_{min}}$
            $p_{i_{min}} \leftarrow p\_value(f_i, f_j)$
            $m \leftarrow j$
    end
    $minimum\_p\_value_i \leftarrow \mathrm{p}_{imin}$
    $k \leftarrow m$
    $NN_{C_0}(f_i) \leftarrow count(\mathrm{f}_j \in kNN(f_i))$ where $f_j \in kNN(f_i)$ is labeled as C_0
    $NN_{C_1}(f_i) \leftarrow count(\mathrm{f}_j \in kNN(f_i))$ where $f_j \in kNN(f_i)$ is labeled as C_1
end
sort $minimum\_p\_value$ in ascending order

Algorithm 2.2 declares the classification algorithm which is based on calculating specific classification score for each subject. The subject's classification score (Γ) is calculated by

Equation 2.19.

$$\Gamma_{subject} = \sum_{i=1}^{n} (p(C_1|f_i) - p(C_0|f_i)) \tag{2.18}$$

$$\Gamma_{feature} = p(C_1|f_i) - p(C_0|f_i) \tag{2.19}$$

where n shows the number of informative feature per subject. C_1 and C_0 illustrate the binary classes. Posterior probability of class given by observed feature, $p(C_j|f_i), j \in \{1, 2\}$ is obtained by counting the number of neighbors belong to the C_j around observed feature, inside the kernel bandwidth.

Algorithm 2.2 Classification

```
for all test Subjects do
    Initialize N_C0 and N_C1 to 0 forall Observed Features f_i do
        N_C0 ← N_C0 + #NN_C0(f_i)
        N_C1 ← N_C1 + #NN_C1(f_i)
    end
    Γ ← N_C1 − N_C0 Subject classification score
end
```

2.5 Parameter Estimation

In the neighborhood space of observed feature, there is a bandwidth kernel density which includes features bearing significant information regarding to the class of observed features. In order to find appropriate bandwidth's size for each observed feature, we seek to estimate the density parameters. Figure 2.5 illustrates the Gaussian distribution of features which is assumed as our density function in this work. Therefore, the variance and the mean of this kernel should be estimated. The 3D location of observed feature, $\bar{x} = (x, y, z)$ is considered as the mean. To estimate the intended variance many statistical significance test may be used, such as: p-value, maximum likelihood estimation, maximum a posterior probability, and mutual information.

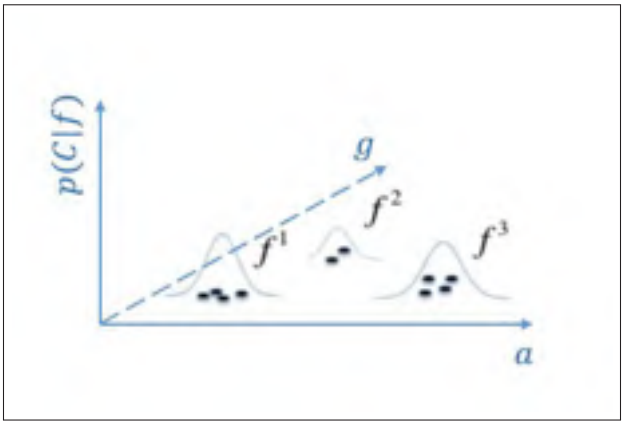

Figure 2.5 Illustrating the distribution of feature data and its neighbors (dots) in appearance and geometry space. In case of large training size, sufficient sampling permits nominally correct correspondence between features arising from the same modes, similar appearance and geometry Taken from Toews & Wells (2016)

2.5.1 p-value

In this statistical model of study a null hypothesis (H_0) is assumed. For example the assumption could be based on that there is a normal curve distribution. The p-value is the probability of the observed data under the null hypothesis (H_0) model of default, uninformative distribution. If the p-value is small, it indicates that the data may not follow an uninformative distribution, and is used as the basis for rejecting H_0, when its amount is smaller than the actual observed results, Wasserstein & Lazar (2016). Over the p-value curve in figure 2.6, the lower amount shows the event was not likely to assume normal distribution which respects the null hypothesis, whereas with a p-value of 1 the null hypothesis is true. In most of studies, scientists consider $p < 0.05$ as a statistically significant which means less than one in a 500 chance of being wrong. The lower p-value obtains the higher confidence in rejecting the null hypothesis.

In this work, we seek to calculate p-value among the nearest neighbors of observed feature f_i. The null hypothesis is assumed that there is a correlation between class labels and features. We did Fisher's exact test as a significance test in order to calculate p-value. Fisher's exact test

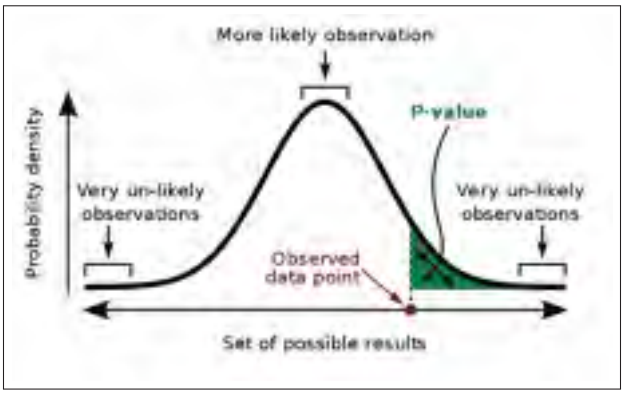

Figure 2.6 The area under the curve over the observed data point, is the p-value
Taken from Wikipedia Wikipedia (2014)
https://en.wikipedia.org/wiki/P-value
Consulted on June 6, 2017

calculates the significance of the deviation from a null hypothesis, Fisher (1992). Fisher proved that hyper-geometric distribution gave us the probability of obtaining any such set of values. In hyper-geometric distribution, probability of k successes in n draws, without replacement, is calculated, Preacher & Briggs (2001). This test is useful in analyzing the nearest neighbors to accurately determine the number of significant nearest neighbors to consider.

Table 2.1 Features in feature-distance space might be considered as the observed features nearest neighbors,$F \in NN$, or observed feature's not-nearest neighbors,$F \notin NN$

	F∈ NN	**F∉NN**
$Class_0$	a	b
$Class_1$	c	d

In table 2.1 each row refers to the corresponding studied class and each column refers to the nearest or non nearest neighbors features. The total number of features in the feature space is $a + b + c + d = n$. Fisher proved that the probability of achieving any such set of values was given by the hyper geometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!(c+d)!(a+d)!(b+c)!}{a!b!c!d!n!} \tag{2.20}$$

In order to compute the p-value the all tables' probabilities given by Equation 2.20 which are less than or equal to the observed table would be added together. In other word, the area under the curve over the observing data point's probability density is the p-value, Fig 2.6.

2.5.2 False Discovery Rate

For this study, it is necessary to find highest informative features per image. The number of most informative features of each subject should be obtained by false discovery rate examination, ϵ_{FDR}, let I be a the test image:

$$p(C|I) = \sum_i p(C, f_i|I) = \sum_i p(f_i|C, I)p(C|I) \simeq \sum_{i=1}^{\epsilon_{FDR}} p(f_i|C, I)p(C|I) \tag{2.21}$$

Each feature has a corresponding p-value number which may be considered as a threshold to find the most informative nearest neighbors. In the image as there are many features there would be many p-values which are corresponded to the features. The p-values are considered as the conducting multiple comparisons and FDR may be used as a method to conceptualize the rate of type 1 errors in this null hypothesis testing.

2.6 Visualization

The goal of this work is to produce a visual overlay for a new patient image $I : R^3 \rightarrow R^1$, e.g. an MRI or CT scan, that can be used to highlight image locations $\bar{x} \in R^3$ linked to clinical variables of interest C, e.g. age, disease, etc.

According to the above points, the intended posterior probability which is considered as an overlay to visualize is $p(C|F, \bar{x})$. We choose posterior probability $p(C|F, \bar{x})$ of the class label C conditional on feature set F and image spatial location to represent the overlay. Using Bayes

rule and assuming conditionally independent feature observations, this may be expressed as:

$$p(C|F,\bar{x}) \propto p(F|C,\bar{x})p(C|\bar{x}) = \prod_{i}^{N} p(f_i|C,\bar{x})p(C|\bar{x}) \propto \prod_{i}^{N} p(f_i|C)p(C) \tag{2.22}$$

where $p(C)$ is a prior probability over the variable of interest and $p(f_i|C)$ is the probability density of an individual feature f_i conditional on the class C. The probability density $p(f_i|C)$ can be factored into conditional densities over appearance a_i and geometry g_i

$$p(f_i|C) = p(a_i, g_i|C) = p(a_i|g_i, C)p(g_i|C) \tag{2.23}$$

and modeled as a kernel density function across a large set of training data, i.e. feature sets F and corresponding variable labels C. In the case where feature f_i is statistically independent of C, we have $p(f_i|C) = p(f_i)$, and there is no class-related information to visualize. We thus calculate the probability of a set of $p(f_i \perp\!\!\!\perp C)$ assuming a null hypothesis of statistical independence. For a feature f_i, we identify a set of k-nearest neighbor features $N_i = \{f_j : kNN(a_i, a_j) \cap kNN(g_i, g_j)\}$ in both appearance and geometry, where f_j is a training feature with associated class label C_j.

Our visualization approach focuses on directly overlaying the probability $p(C|f_i,\bar{x})$ of a class C conditioned on individual observed feature information f_i, e.g. geometry/location and appearance. This allows a human expert to directly visualize image locations $\bar{x}_i$ where anatomical tissue patterns are informative regarding the value of C. This probability is related to the density $p(f_i|C)$ in Equation 2.23 via Bayes rule as follows:

$$p(C|f_i) = \frac{p(f_i|C)p(C)}{p(f_i)} \propto p(f_i|C)p(C) \tag{2.24}$$

where $p(f_i|C)$ may be generally modeled using KDE. A key challenge is estimating the crucial kernel bandwidth parameter, which generally varies across feature space and with the number of data samples in memory.

CHAPTER 3

EXPERIMENTS

3.1 Data and Features

The brain MRI dataset is the focus of our experiments. Data is being collected from the Open Access Series of Imaging Studies, OASIS Marcus *et al.* (2007) which is series of publicly available neuroimaging datasets. This dataset which is acquired to aim future discoveries in clinical neuroscience, is examined in many terms such as age and gender, in this experiment. Also, this experiment focuses on analyzing Alzheimer's Disease (AD's), one of the important health problem of these days. Table 3.1 shows the demographic characteristics of this study. This dataset contains 416 subject's cross-sectional brain MRI collection. One hundred of them suffers from Alzheimer's, with very mild to moderate level. Age range of the subjects varies between 18 to 96 years old. The dataset includes right-handed of both men and women. OASIS provide three or four individual T1-weighted MRI scans in a single image for each patient. For the reason that these data are accommodatable to a wide range of analytic approaches such as automated computational analysis, we used this dataset in our research. Owing to this fact that there is a large amount of information inside of the images, it is more efficient to focus on the salient and interesting features.

Table 3.1 Demographic characteristics of the study groups

	Number of subjects	**#Male**	**#Female**	**Age(Average ± StDev)**
OASIS dataset	416	160	256	52±25

In the figure 3.1 the structure of salient features extraction is shown by the diagrams. The first step before extracting the features is to align all images in to the standard frame Talairach & Tournoux (1988). In this work the images are registered to the Talairach standard atlas. Then SIFT features are extracted from the images and stored to the key files. The name of each file is the number of each patient in the list of the OASIS dataset. Also, the second output is a csv file which

contains all the details about each patient, such as their gender, age, CDR and many other clinical factors. CDR or the Clinical Dementia Rating is a numeric scale used to quantify the extremity of symptoms of dementia. This clinical factor is essential nowadays as the dimentia-related illness such as Alzheimer's Disease have been pervasive. The theory of feature extraction step is described in section 2.1.

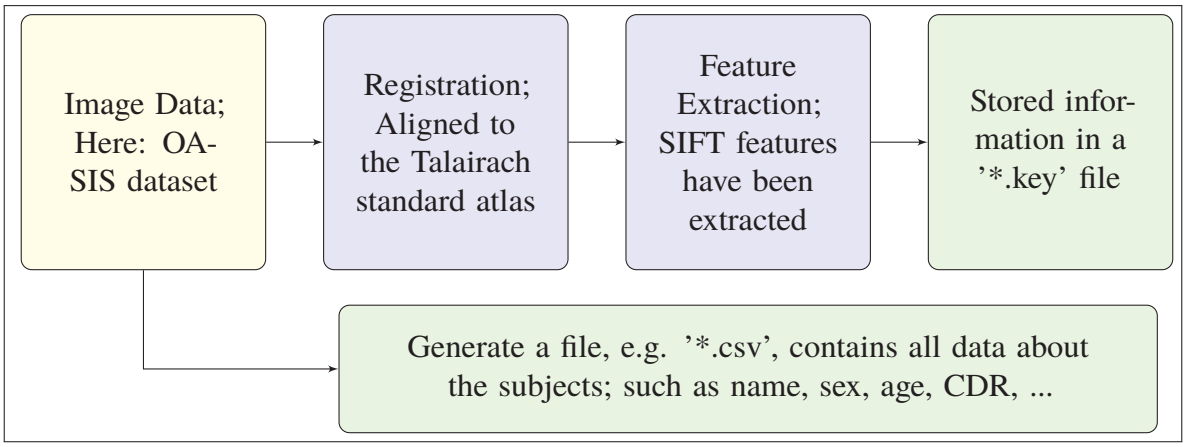

Figure 3.1 Data pre-processing

The initial data which have been used in this experiment are SIFT features extracted from each patient brain's MRI. These features contains local coordinate system descriptors, $\{X, \sigma, \Theta\}$. Where the three location parameters which specified by the origin are shown by $X = \{x, y, z\}$, scale is noted by σ. $\Theta = \{\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\}$ is the set of three orthonormal unit vectors specifying the rotations of coordinate axes which is not utilized in this experiment. Also these key files include 64 appearance descriptors for each salient feature, Toews & Wells (2013).

The first step in data analysis is data pre-processing,and preparation such as data labeling, labeling structure, augmentation, and feature extraction, to improve the dataset quality. Figure 3.2 shows labeling structure in this study. One of the clearest reason to have a low quality data is that it may consist of many uninformative data. Recognizing and removing uninformative data would improve the quality of dataset. In order to find uninformative data in this dataset (SIFT features of OASIS brain's MRI), we need to see the features in a feature space. Then comparison would be able. Seeing high dimensional data is a big challenge as it could not be visualized

in 2D or 3D space. There are many ways to see them such as using colors, different shapes. Nevertheless if the dimension is more than 6, the visualization might be hard to understand. Feature-distance space is a good idea to visualize and see differences between the features in this dataset as geometric and appearance variables are conceptually relative to their specified origin points.

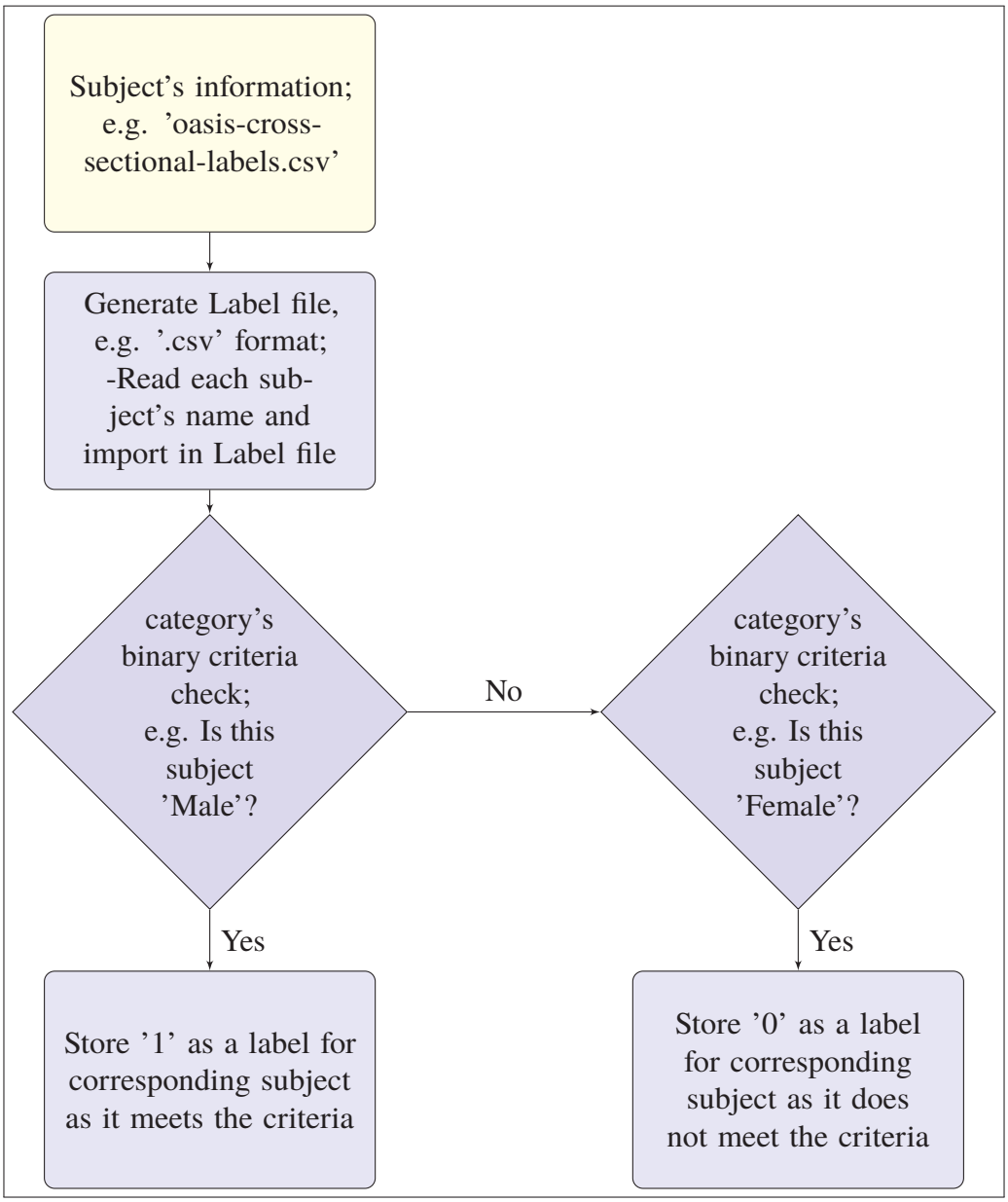

Figure 3.2 One Sample of Given Subjects Labeling Structure

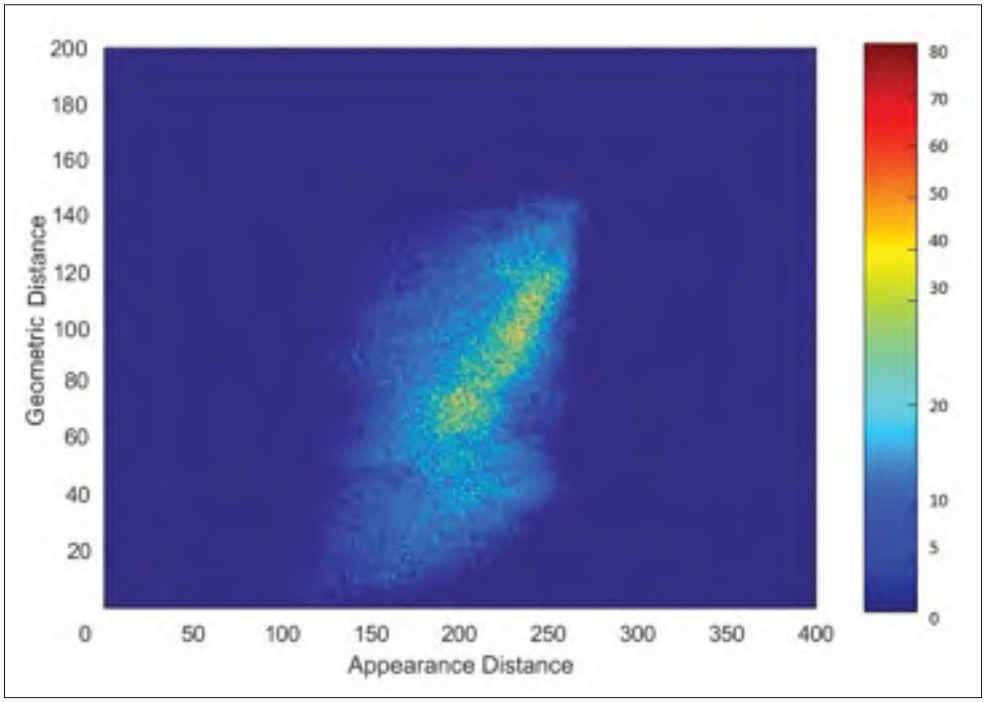

Figure 3.3 Example distance distribution betweeen one feature and all other features in the dataset, where distance is visualized here in independent axes of geometry (vertical) and appearance (horizontal), distances are not yet normalized

Figure 3.3 is provided to show high dimensional features in 2D space. It illustrates histogram of euclidean distance between the geometry and appearance descriptors between of one sample of observed feature coming from a test subject and other feature in the whole feature dataset. It is shown that there are many features inside of the big cloud that do not have any statistical linear relationship. With a closer look, there are a few features close to the origin of the histogram that may be more informative.

The goal of this experiment is to see the similarities and differences between different subject's brain images. In medical data analysis, the clinical researchers may consider the appearance differences more important rather than the geometry differences. There is a fact that with high differences between geometry descriptors, the anatomical location would refer to different part or anatomy. Most of the time, the clinical researchers are looking to find differences and similarities between same anatomies among the different subjects. Therefore, in the feature-distance space

the cloud of data and the features with high geometric distance are not considered as being informative in this experiment. Consider that in this dataset, brains images are aligned to a standard frame and the notation of (x, y, z) in all the images indicate to the same position in all the brains.

3.2 Classification

The k-nearest neighbour kernel density estimation method is applied to the selected subsets to investigate image classification. In this research, three different subsets of the OASIS dataset are tested to validate our method, shown by tables 3.2, 3.3, and 3.4. For gender category, the subjects who are less than 60 years old are considered. For age category, the threshold to determine being old or young is calculated by the median of this dataset's age range which is 57. And for Alzheimer's disease, clinical dementia rating (CDR) is considered to categorise subjects. This subset contains the subjects who are more than 60 years old.

Table 3.2 Gender category's demographic table

Gender	**Subjects**	**Age(Average ± StDev)**
Female	119	32±13
Male	119	30±12

Table 3.3 Age category's demographic table

Age	**Subjects**	**Male**	**Female**	**Age(Average ± StDev)**
Old	208	72	136	75±8.8
Young	208	88	120	30±12

Table 3.4 Disease category's demographic and clinical characteristic table

Disease	**Subjects**	**Male**	**Female**	**Age(Average ± StDev)**	**Clinical Dementia Rating**
Alzheimer's	97	41	59	76.76±7.12	0.5; 1; 2
Healthy	97	26	72	75.92±8.98	0

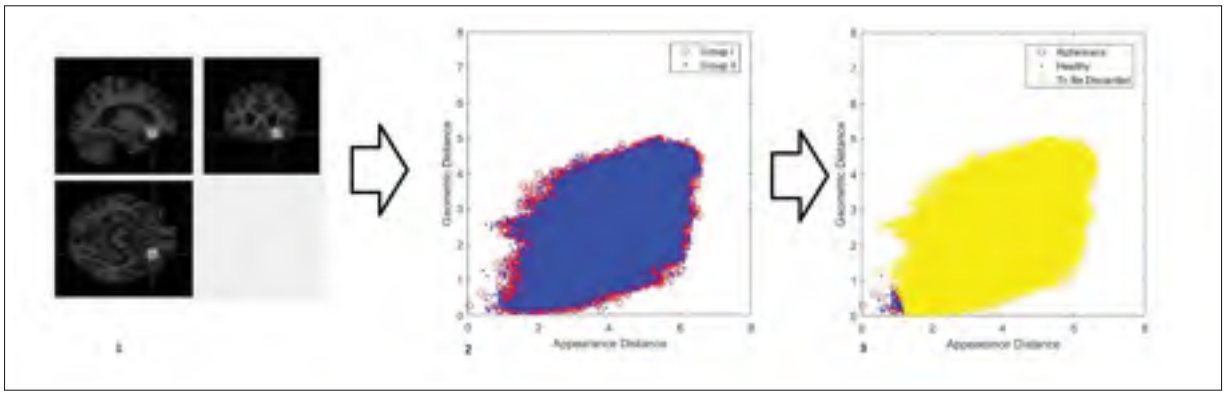

Figure 3.4 Finding informative nearest neighbors; From left to right: 1) Select a feature from the subject's MR image (observed feature), 2) Sort all the other features in the feature space based on their distance to the observed feature, 3) discard the cloud of features which are not informative

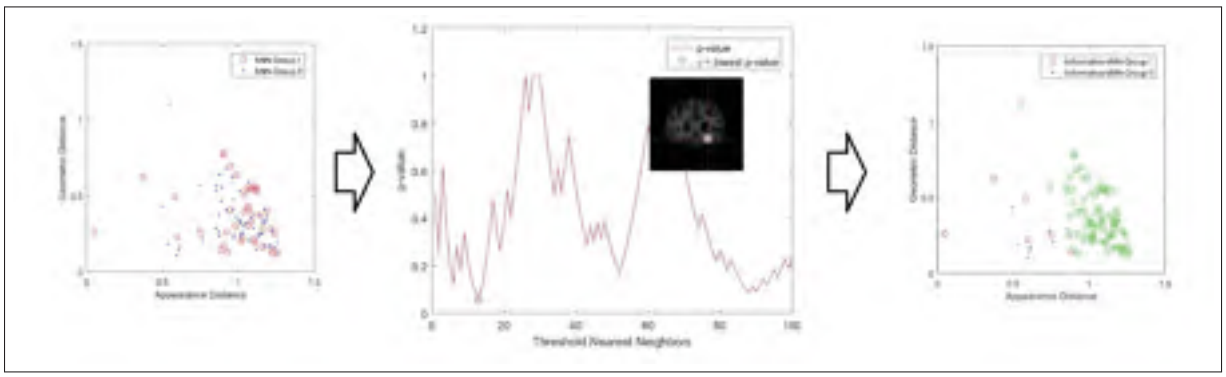

Figure 3.5 Calculate p-value among the nearest neighbors of observed features, index of the feature corresponded to the lowest p-value illustrates the number of most informative nearest neighbors

In this study, each subject has approximately $1k$ salient features. There is more than 100k features in the whole dataset that most of them are too far to be considered as the informative neighbors of observed feature in the classification model. The procedure of discovering informative nearest neighbors for each feature of the subject is shown by figure 3.4 and 3.5. The procedure of finding the most informative nearest neighbors is done by the following steps: First is to normalize and sort the observed feature's neighbors in geometry and appearance descriptors based on their distance metrics. Note that only k-nearest neighbor are considered as informative features around

observed feature to discard the intense cloud of features which are not informative. Second is to calculate and draw graph of p-value for k-nearest neighbors, and third is to find the lowest p-value to reject the null hypothesis of being non-informative. By rejecting the null hypothesis, most informative nearest neighbors are discovered. The procedure of training step has been shown in figure 3.6.

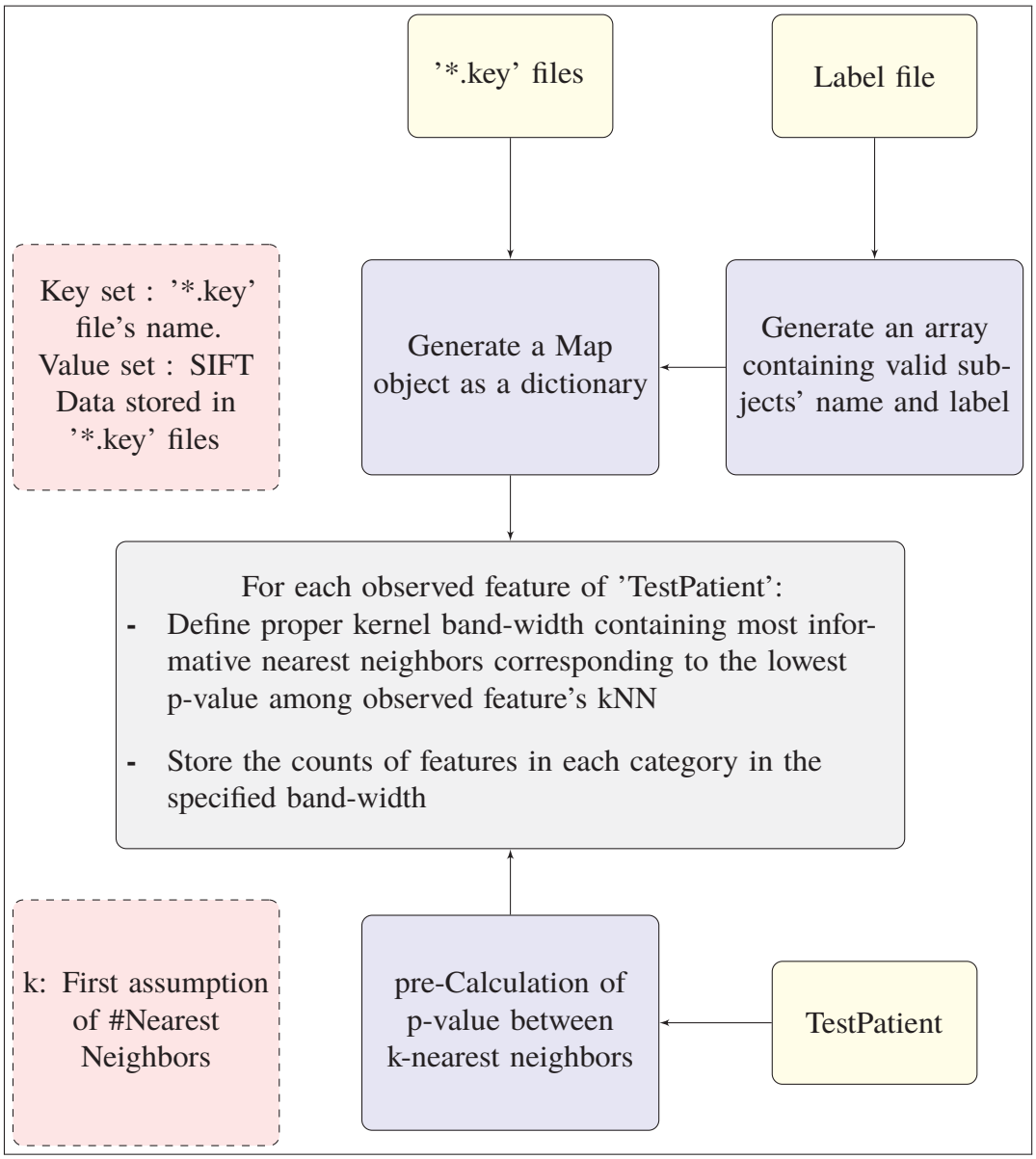

Figure 3.6 Main program structure untill training section

A Map object is a data structure that allows you to retrieve values using a corresponding key. It provides more flexibility for data access than array indices, which must be positive integers.

In this study, hypothesis test aims to find the proper bandwidth of the kernel in order to classify the subjects, figure 3.7. This figure shows that around each observed feature, there are many features belong to the different classes. The kernels may have different size of band-widths. It is necessary to find the most informative band-with regarding to the most informative number of nearest neighbors.

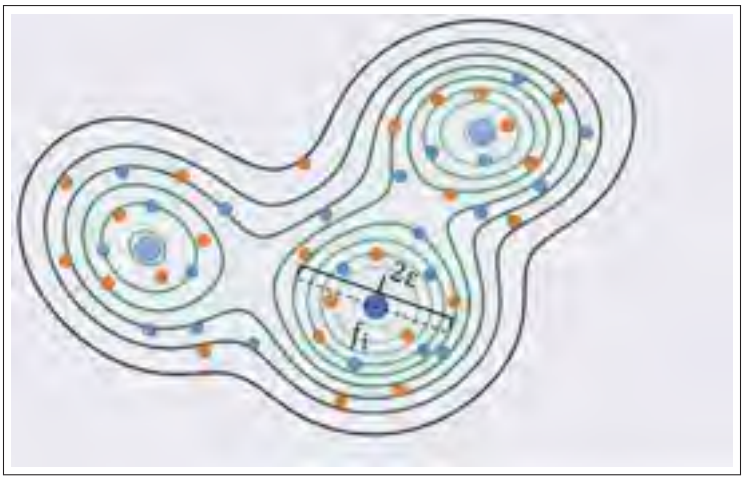

Figure 3.7 Kernel Density Estimation's band-width

By defining proper kernel's band-width, the most informative nearest neighbors of observed feature is found. The number of features belong to each class around observed feature is shown by $p(C_j|f_i)$, where $C_j \in \{C_1, C_2\}$ refers to the binary classes and f_i is observed feature. Difference of posterior probability of class given by observed feature,$p(C_1|f_i) - p(C_2|f_i)$ may get the informative characteristic value of the observed feature. We have to consider that each subject contains many observed feature. Therefore, their characteristic value have to be added together to achieve the classification score per subject, Γ equation 3.1. $\Gamma_{subject}$ is calculated by

the following equation which is explained in section 2.4:

$$\Gamma_{subject} = \sum_{i=1}^{n} (p(C_1|f_i) - p(C_0|f_i)) \tag{3.1}$$

In order to calculate the classification score of each subject, we may have to consider the informative observed feature of each image. False discovery rate would aim to find the most informative feature per subject. FDR shows the rate of type I errors testing when conducting multiple comparisons. In this experiment, we set α, $FDR = p(H_0|p \leq \alpha)$, according to the best classification accuracy result, by calculating the classification accuracy rate based on different number of features per subject. Figure 3.8 illustrates the classification accuracy results versus the different number of features per subject for age dataset. Here, 216 is chosen as the best informative features per subject.

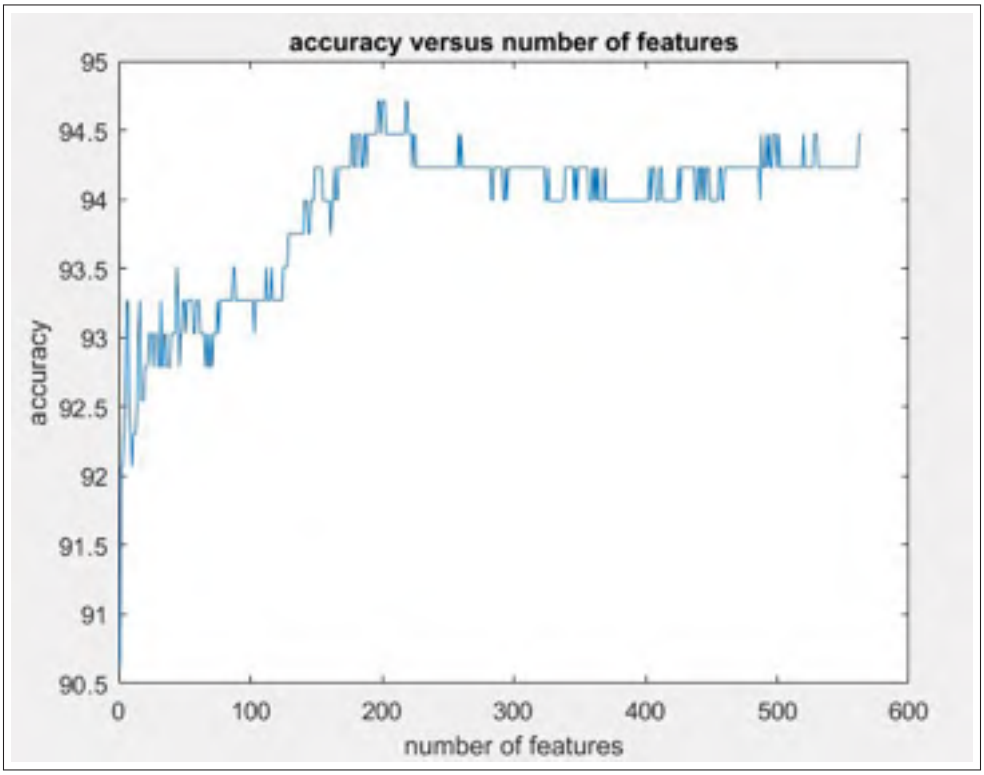

Figure 3.8 Classification accuracy rate vs. number of features per subject

Figure 3.9 shows a line with the slope equal to the achieved alpha which has an intersection with the graph of sorted p-values for one random subject.

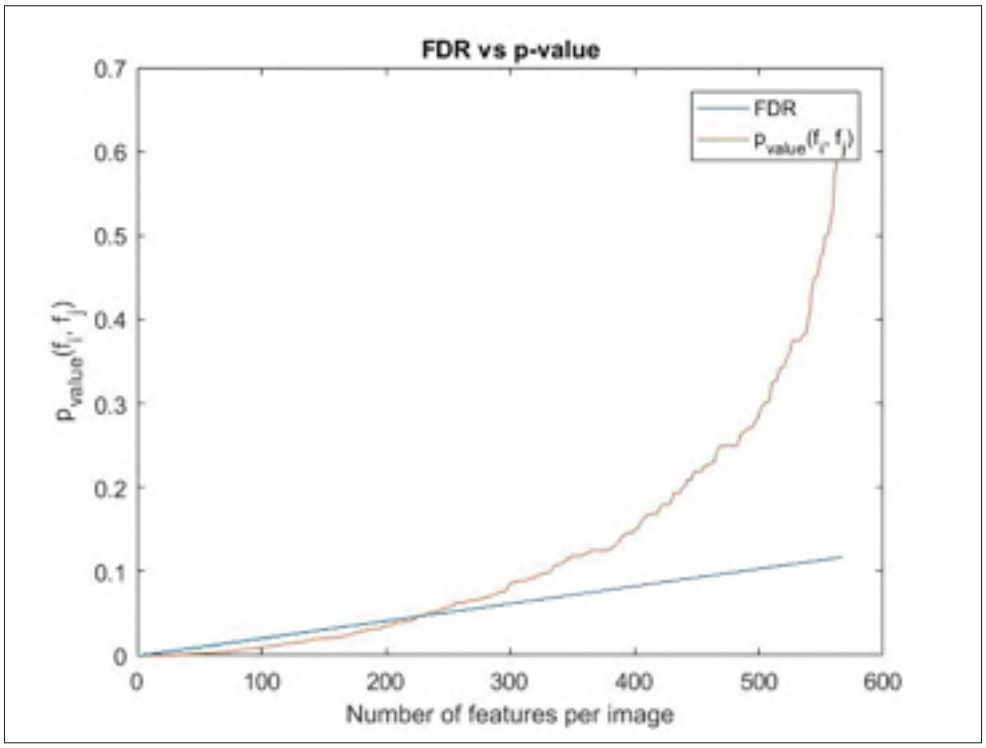

Figure 3.9 FDR vs. features $\{f_i\}$ from a single subject MRI, sorted in increasing order of minimum p-value,$\alpha = 0.05$

Figure 3.10 shows three diagrams including the subjects and their calculated classification scores,Γ, in sorting order.

Figure 3.11 shows the histogram of subject's classification score for age dataset which are nearly perfectly separated. Classification score,Γ is the characteristic value of each subject depends on the difference of posterior probability of class ($C = \{c_1, c_2\}$) given by the observed feature, $\Gamma = \sum_i p(C = c_1|f_i) - p(C = c_2|f_i)$. The intersection between distributions is considered as the threshold for classification.

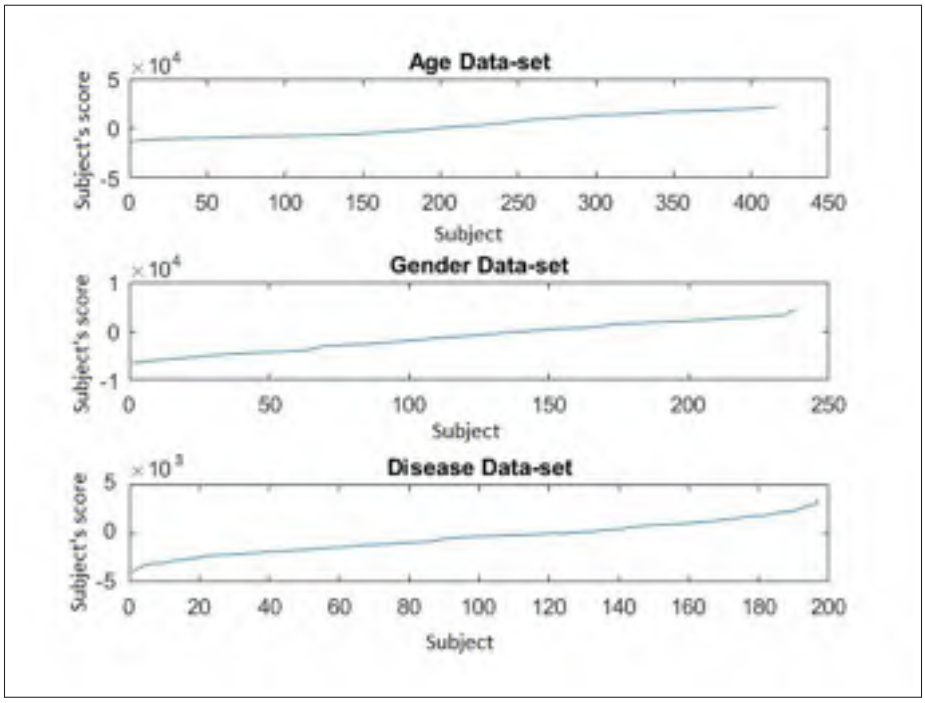

Figure 3.10 Sorted subject's classification scores are shown by this diagram in three different graphs relate to different subsets

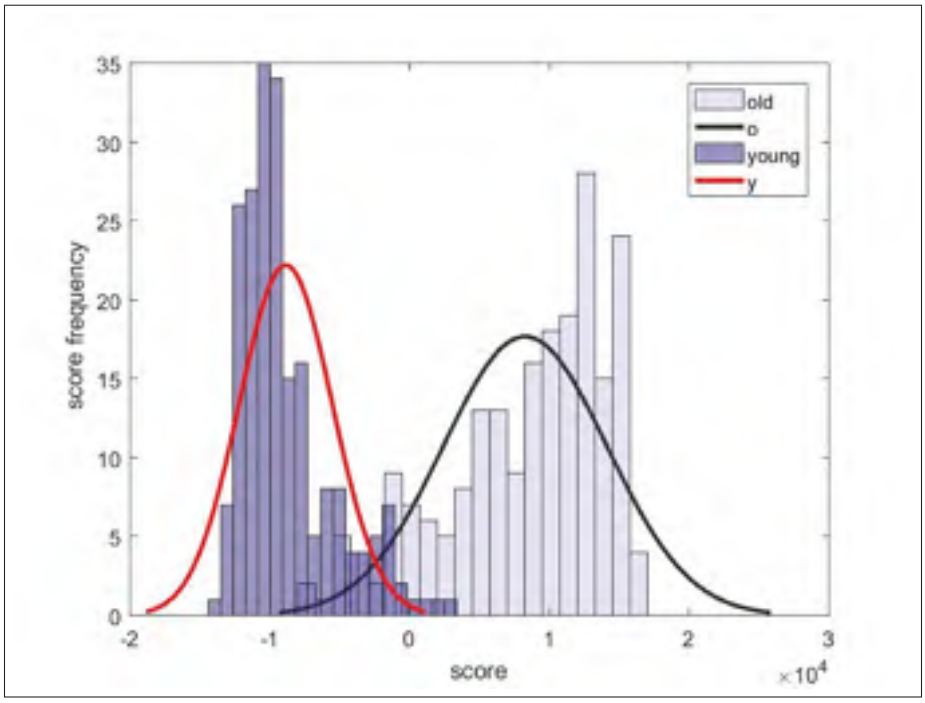

Figure 3.11 Distribution of classification score(Γ); Age subset

Figure 3.12 shows ROC curve of classification for different subsets of OASIS dataset. Blue line refers to the age group with classification accuracy rate 94.71%, red line relates to the gender group with classification accuracy rate 71.2%, yellow line relates to the disease group with classification accuracy rate 70.1%. Age group size is 416 containing 208 old (with the age more than 57 years old) and 208 young subjects (with the age less than 57 years old). Gender group size is 238 (all of the subjects are less than 60 years old), containing 119 males and 119 females. Disease group contains 194 subjects with the age more than 60 years old, 97 Alzheimer's and 97 healthy subjects. Figure 3.13 illustrates different training dataset size versus corresponding classification accuracy rate. The classification accuracy increase slightly in proportion of the size.

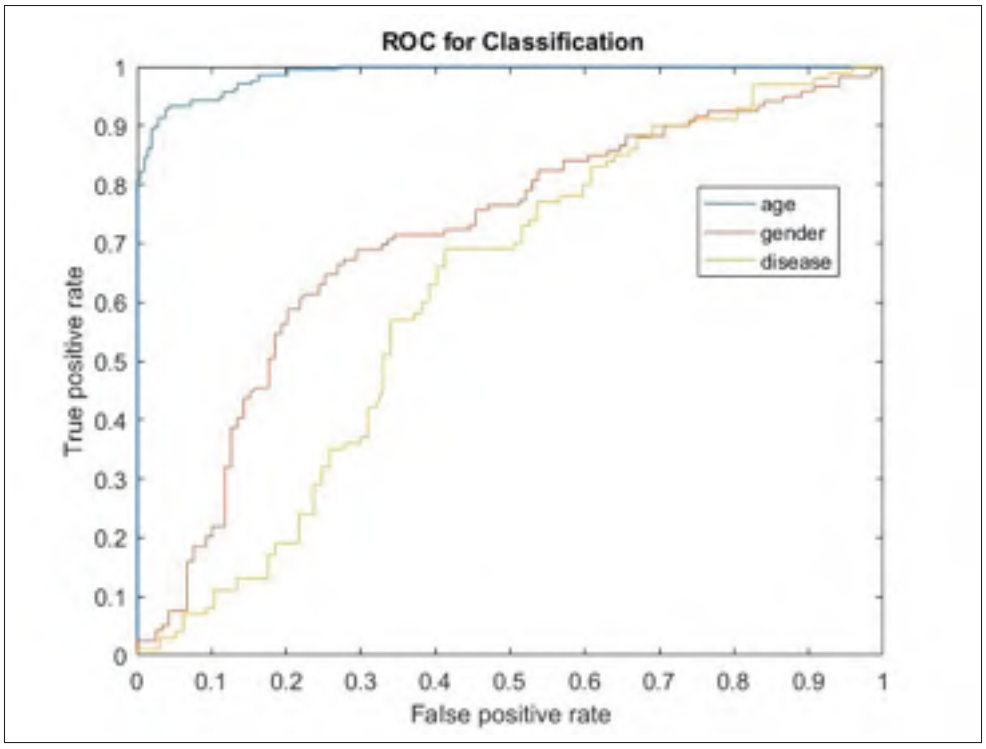

Figure 3.12 ROC classification curve; Blue line refers to the age group with classification accuracy rate 94.71%, red line relates to the gender group with classification accuracy rate 71.2%, yellow line relates to the disease group with classification accuracy rate 70.1%

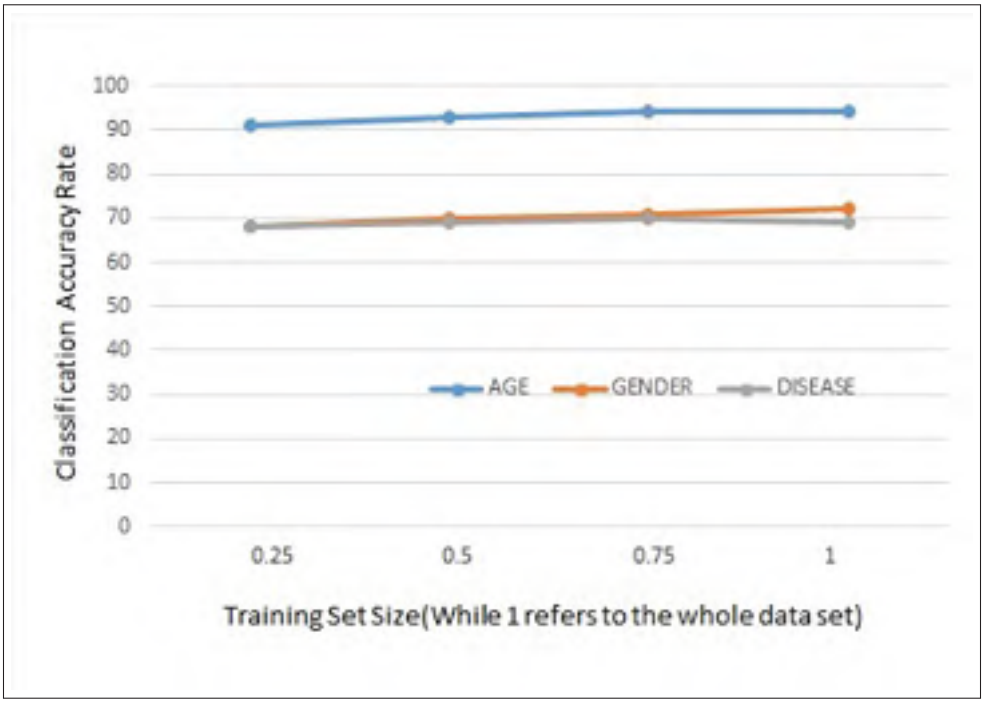

Figure 3.13 Accuracy of brain MRI classification vs. training set size; the classification accuracy increase slightly in proportion of the size

Table 3.5 shows classification accuracy results for all subsets.

Table 3.5 Accuracy of brain MRI classification on the different subsets of OASIS dataset

dataset	Accuracy	Recall	Precision
Age	94.71%	91.82%	97.44%
Gender	71.2%	72.32%	68.07%
Disease	70.10%	68.76%	71.13%

Table 3.6 shows classification accuracy results for all subsets with the binary category's details.

Table 3.6 Detailed classification accuracy rate

dataset	Category	Accuracy
Age	Old	91.83%
	Young	97.59%
Gender	Male	73.95%
	Female	68.06%
Disease	Alzheimer's	69.03%
	Healthy	71.13%

3.3 Visualization

The main goal of this experiment is to visualize the highly informative features to help expert human to see and interpret the images faster and easier. In order to visualize the informative features, we generate an overlay for each background image which contains the posterior probability of class given by feature and spatial location, $p(C|F,\bar{x})$. Features are extracted from a group of subject's brain MRI or individual subject's image. To visualize the medical data coming from a group of subjects with an interest label, group-related feature heat map overlay is generated. It is interested to see the medical data related to image-wise. Group-wise and image-wise visualization are generated for three different subsets of subject's brain MRIs. Figure 3.14 shows a sample of brain anatomy in three planes, sagittal, coronal, and axial. To see the name and location of brain's anatomy, a normal anatomy atlas is generated by the Whole Brain Atlas, an online resource for central nervous system imaging developed by Keith Johnson, and Alex Becker[1].

[1] Med.harvard.edu.(2019).
Available at: http://www.med.harvard.edu/AANLIB/cases/caseNA/pb9.htm.

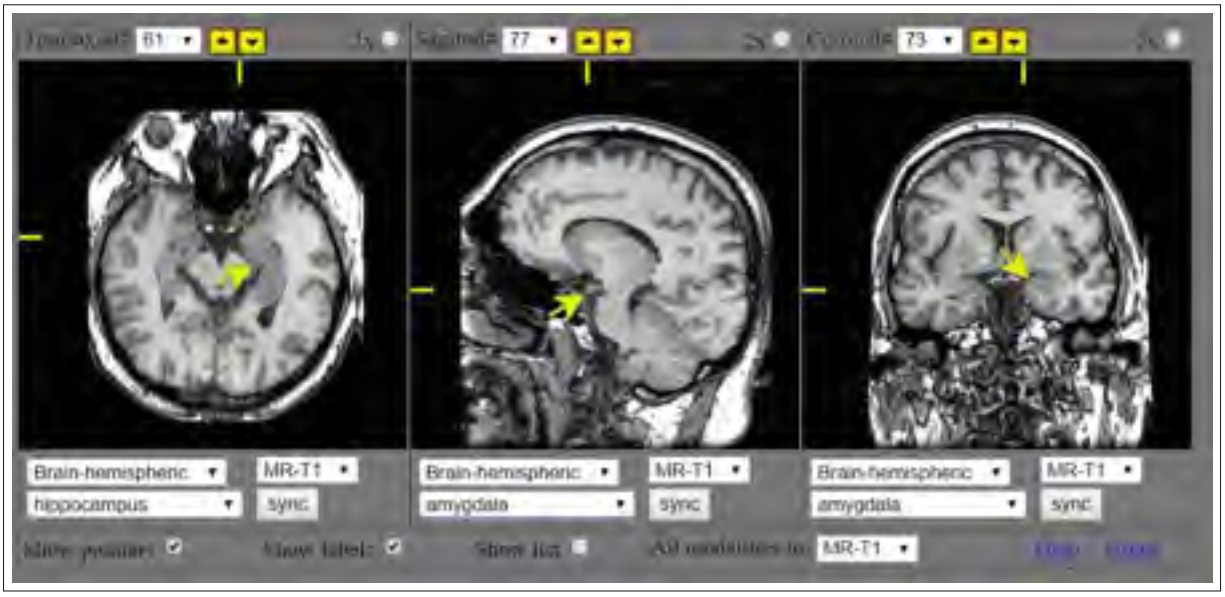

Figure 3.14 A sample of brain anatomy shown by normal anatomy in 3-D with MRI/PET

An example of software platform to have three-dimensional visualization of brain MRI is 3D Slicer, an open source software platform for medical image informatics, image processing. Distribution of informative features are displayed as heat maps (bright pixels) overlaying 3D anatomical MRIs in sagittal, axial and coronal planes and in a 3D rendering by figure 3.15.

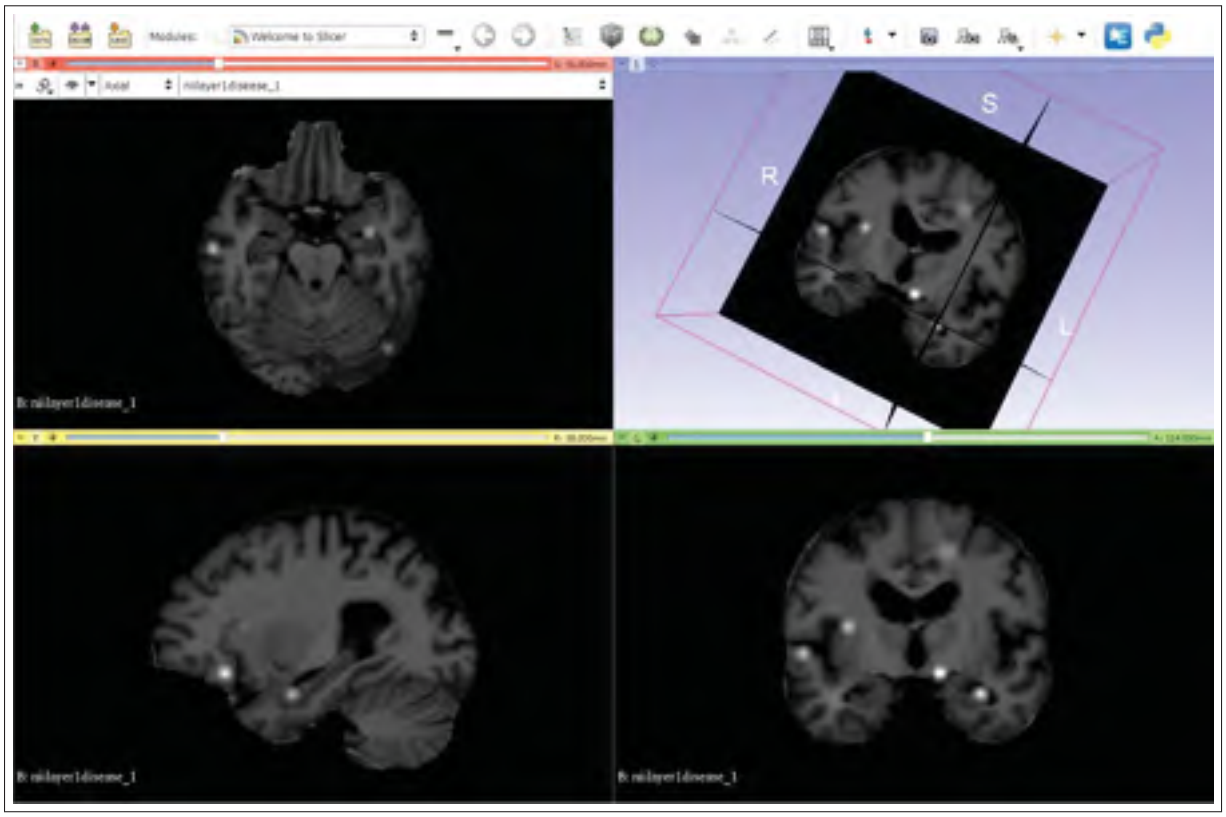

Figure 3.15 An example visualization of a feature distribution using 3D Slicer, an open source software platform for medical image informatics, image processing, and three-dimensional visualization. Distribution of informative features are displayed as heat maps (bright pixels) overlaying 3D anatomical MRIs in sagittal, axial and coronal planes and in a 3D rendering (upper right image)

3.3.1 Group-wise Visualization

To visualize posterior probability of group label given group-related features, heat map overlay is generated. In medical imaging, this topic would be attractive and important as it could help medical scientist to discover the disease's features by seeing to the images.

Figure 3.16 shows the group-wise Visualizing distributions of image feature data F vs. subject categories C. The upper two graphs represent distributions of probability classification scores for all MRI images. The lower images display the posterior probabilities $p(C|F,\bar{x})$ of group label C conditioned on feature set F and spatial image location $\bar{x}$. On the left, blue and red

represent feature distributions for young and old age categories. On the right, the same colours represent healthy and Alzheimer's disease categories. The interesting point of this experiment is seeing symmetric characteristic in age and disease groups which also point to similar anatomies in old and also disease groups about the mid-sagittal plane. In this figure, blue circles refer to the old and disease people, who are more than 57 and more than 60 years old, respectively. 57 years old is the median age of dataset and considered as a threshold for being old or young. All the disease subjects are more than 60 years old. Apparently, they may have common significant features. Red circles refer to young and healthy subjects who are less than 57 and more than 60 years old, respectively. therefore, we do not expect to see the features in the same anatomies.

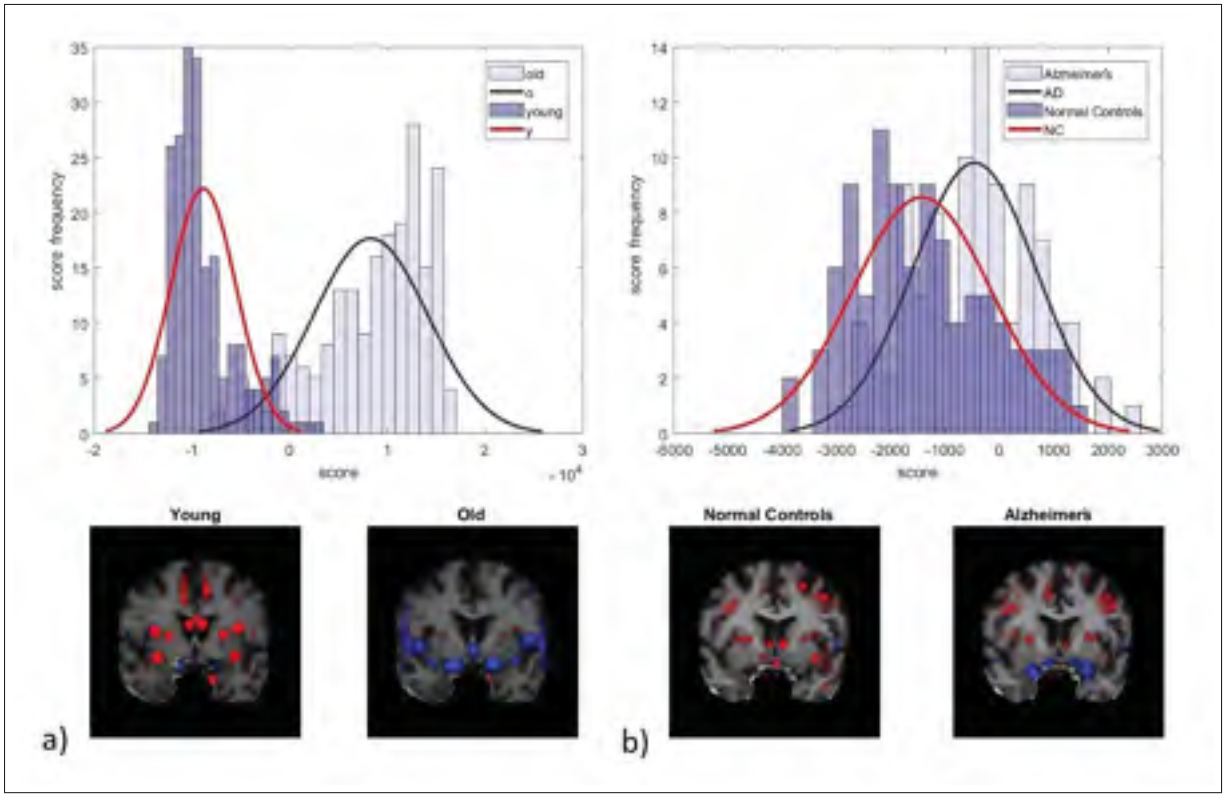

Figure 3.16 Visualizing distributions of image feature data F vs. subject categories C. The upper two graphs represent distributions of probability classification scores for all MRI images. The lower images display the posterior probabilities $p(C|F,\bar{x})$ of group label C conditioned on feature set F and spatial image location $\bar{x}$. On the left, blue and red represent feature distributions for young and old age categories. On the right, the same colours represent healthy and Alzheimer's disease categories

Figure 3.17 shows group-wise visualization of the posterior probability $p(C|F,\bar{x})$ of group label C conditioned on feature set F and spatial image location $\bar{x}$. On the left, red region represent the sagittal and axial plane's feature distributions for young group. On the right, the blue colours represent feature distributions for old group. The interesting part of this visualization is seeing symmetric characteristic about the mid-sagittal plane in left and right hemispheres and also most of the features are concentrated in the basal ganglia and cerebral cortex.

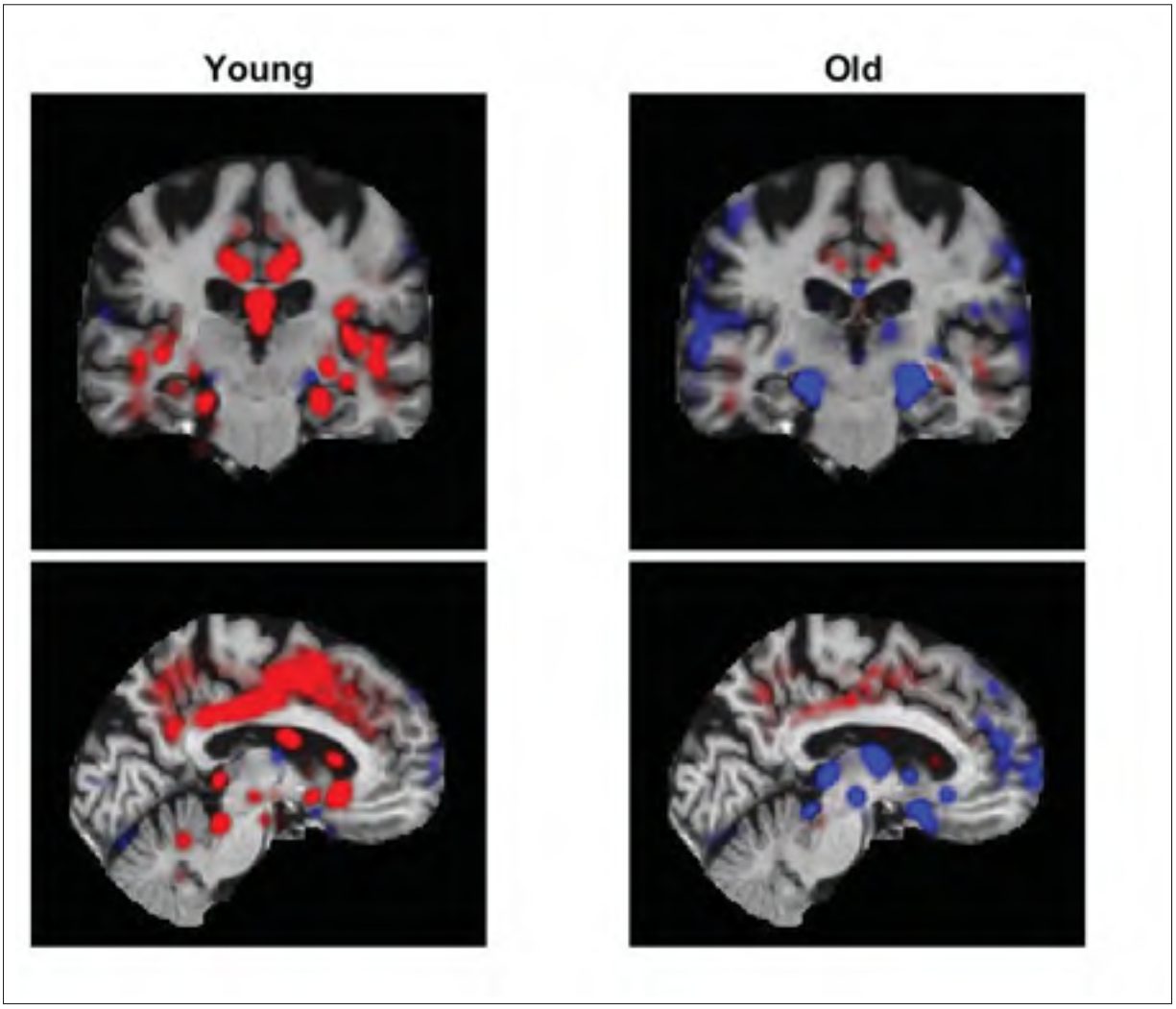

Figure 3.17 Group-wise visualization of the posterior probability $p(C|F,\bar{x})$ of group label C conditioned on feature set F and spatial image location $\bar{x}$. On the left, red region represent the sagittal and axial plane's feature distributions for young group. On the right, the blue colours represent feature distributions for old group

Figure 3.18 shows group-wise visualization. The heat map overlay shows the posterior probability $p(C|F,\bar{x})$ where $C \in \{age, disease, gender\}$. The probability intensity changes are shown by the using heat map. Higher intensity shows higher $p(C|F,\bar{x})$. The top image relates to the age group. There is symmetric characteristics about mid-sagittal plane. Most of the features are concentrated about basal ganglia. The middle one relates to the disease group. We can see symmetric characteristic here also. The lower image is showing gender group which has

symmetric and asymmetric features about mid-sagittal plane and the features are concentrated in basal ganglia and also cortex.

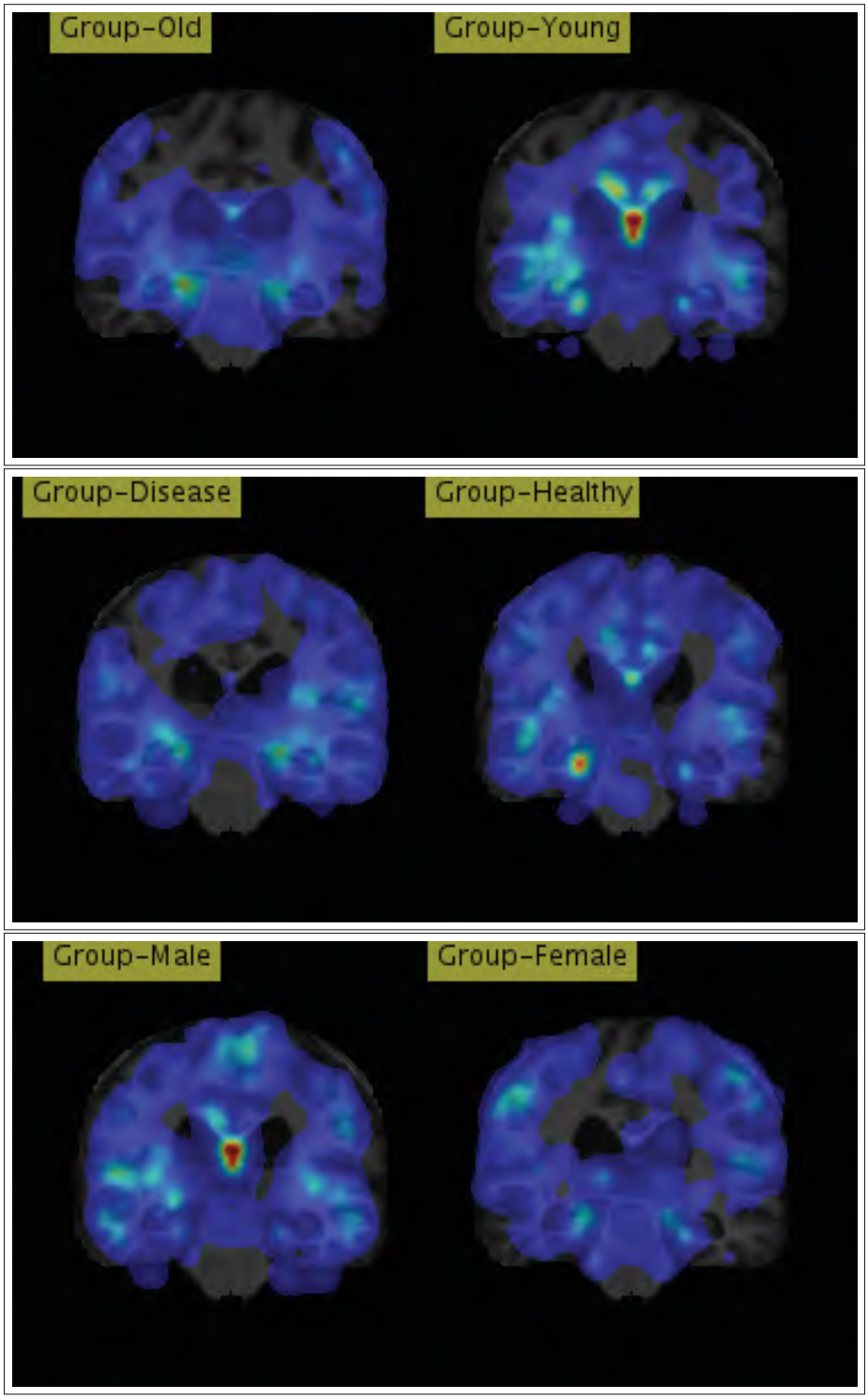

Figure 3.18 Group-wise Visualization of different sub-datasets;heat map overlay on the same coronal plane

3.3.2 Image-wise Visualization

Image-wise visualization is a technique to visualize the posterior probability $p(C = "label"|\bar{x}, F)$ where F is a set of features coming from individual subject. This type of visualization is helpful to diagnosis the disease symptoms, find similarities between same anatomies between different subjects and etc. in medical imaging.

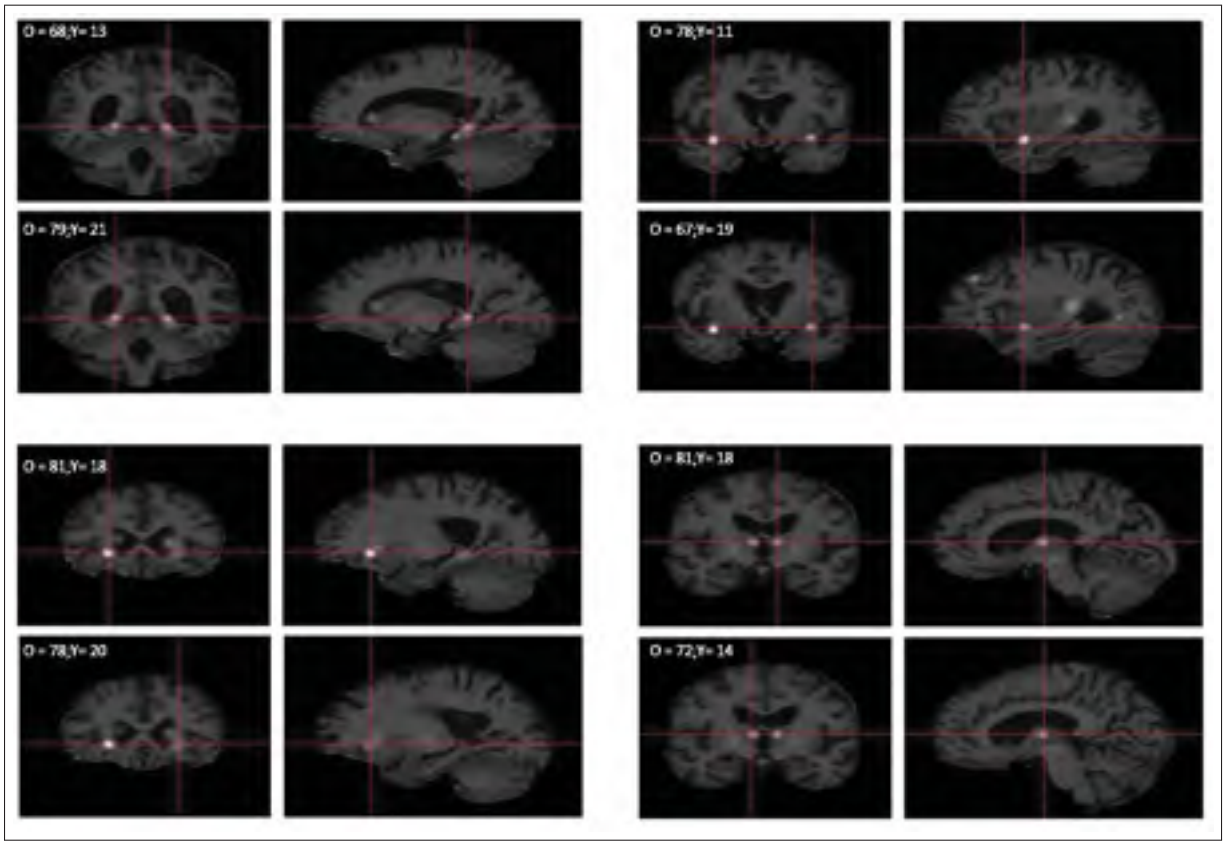

Figure 3.19 Individual feature visualization of $p(C = Old|F, \bar{x})$ for a healthy (CDR=0) elderly male individual (age=87), in different slice planes. Note that age-informative features are generally concentrated the left and right hemispheres in anatomical locations that are symmetric about the mid-sagittal plane

Figure 3.19 shows individual feature visualization of $p(C = Old|F, \bar{x})$ for a healthy (CDR=0) elderly male individual (age=87), in different slice planes. Note that age-informative features are generally concentrated the left and right hemispheres in anatomical locations that are symmetric about the mid-sagittal plane.

Figure 3.20 and 3.21 shows feature visualization of $p(C = Old|F, \bar{x})$ for a healthy (CDR=0) elderly male individual (age=87). The visualized features are the most significant features (with the lowest p-value score) among the subject's features and also the feature of left hemispheres is symmetrical with the right one about the mid-sagittal plane. Corresponding to the p-value score.

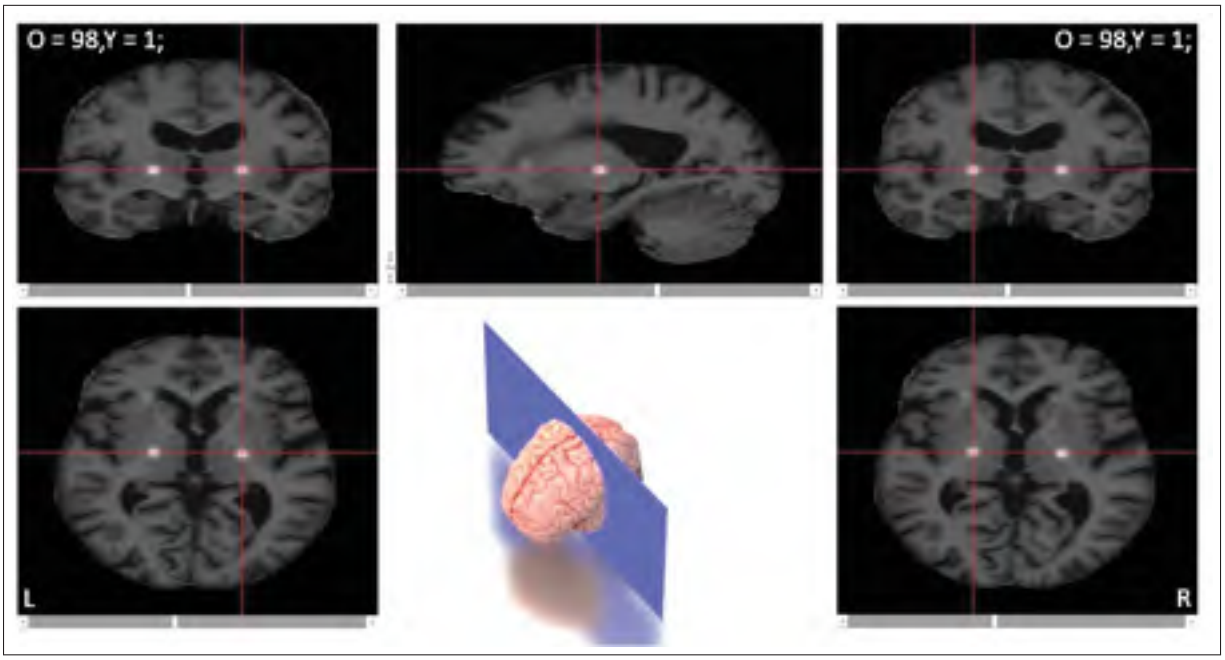

Figure 3.20 Individual feature visualization of $p(C = Old|F, \bar{x})$ for a healthy (CDR=0) elderly male individual (age=87). The visualized feature is the most significant feature (with the lowest p-value score) among the subjects features and also the feature of left hemispheres is symmetrical with the right one about the mid-sagittal plane. Corresponding to the p-value score($\propto 10^{-28}$), k is set to 99 containing only one feature from young category for both side

To conclude, the principal contribution of this work is to visualize significance features which aims to automatically identify subject's age range, gender, and also discriminate between healthy subjects and diseased subjects with Alzheimer, by their whole-brain magnetic resonance images, MRI. MRI is a popular high-quality medical imaging technique without any bad radiation effect on the body. On the other hand, MRI is the best choice to distinguish tissue characteristic differences easily. To visualize informative data on the MRI would help the clinicians and researchers to interpret the images faster and easier.

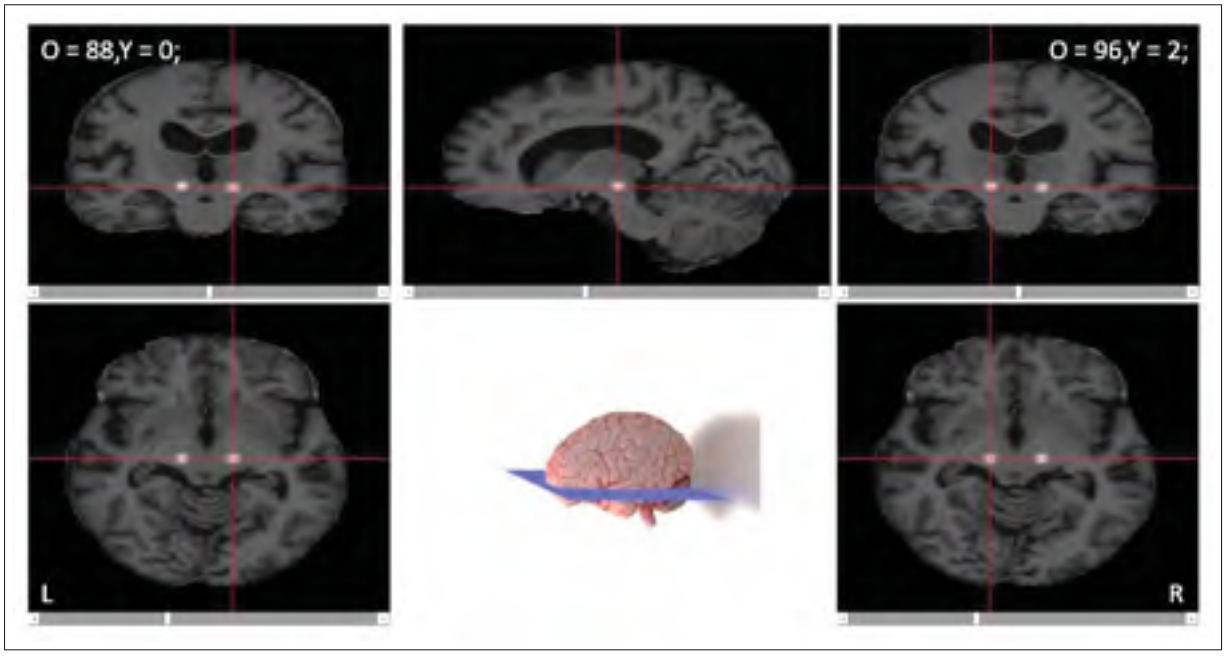

Figure 3.21 Individual feature visualization of $p(C = Old|F, \bar{x})$ for a healthy (CDR=0) elderly male individual (age=87). The visualized feature is the most significant feature (with the lowest p-value score) among the subjects features and also the feature of left hemispheres is symmetrical with the right one about the mid-sagittal plane. corresponding to the p-value score($\propto 10^{-27}$), k is set to 88 without any young feature for the left side; however, right side has 98 nearest neighbors with 96 coming form elder group

CONCLUSION AND RECOMMENDATIONS

This study analyses and proposes a data visualization paradigm for human interpretation, visualizes the summary of highly-informative class-related information obtained by kNN kernel density estimation. We provide a survey of feature-based classification and visualization that it might have a broad range of applications in medical and non-medical scientific data analysis in many disciplines. OASIS brain MRI data was analyzed as our dataset in this experiment.

In experiments using brain MRI data, selecting the bandwidth of kernel density by considering the statistical analysis and framework on the neighborhood marginalization of observed scale invariant feature transform, result in the brain classification rate of age sub-dataset is reported for the OASIS dataset, with an AUC=94.71%.

Experiments in visualization of informative features use heat map overlay of the posterior probability of class given by the spatial location and feature dataset,$p(C|\bar{x}, F)$. We demonstrate there are many symmetric features about mid-sagittal plane in the left and right side of brain hemispheres of age and disease subsets. Also there are many asymmetric features shown in the results for gender subsets. Also, informative feature visualization accelerates the understanding and interpreting the images.

Future work will involve further investigation on multi-classes regression and classification rather than only binary classification, e.g. to predict exact age of the subject instead of just predicting the age range, old and young. In addition, this classification and visualization paradigm could be examined on various datasets such as non-medical images and also other medical datasets. Also defining a proper bandwidth by examining other methods such as maximum likelihood estimation, maximum a posterior probability, mutual information would be interesting.

REFERENCES

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3, 103–107.

Abid, A. M. . (2013). Introduction to SIFT (Scale-Invariant Feature Transform). Consulted at http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html.

Allaire, S., Kim, J. J., Breen, S. L., Jaffray, D. A. & Pekar, V. (2008). Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1–8.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.

Amit, Y. & Kong, A. (1996). Graphical templates for model registration. *IEEE Transactions on pattern analysis and machine intelligence*, 18(3), 225–236.

Ashburner, J. & Friston, K. (1997). Multimodal Image Coregistration and Partitioning - a Unified Framework. *NeuroImage*, 6(3), 209–217.

Ashburner, J. & Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, 11(6), 805–821.

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J. & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial psychiatry journal*, 18(2), 127.

Bay, H., Tuytelaars, T. & Van Gool, L. (2006). Surf: Speeded up robust features. *European conference on computer vision*, pp. 404–417.

Bender, R. & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4), 343–349.

Bengio, Y. et al. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1), 1–127.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.

Bersvendsen, J., Toews, M., Danudibroto, A., Wells III, W. M., Urheim, S., Estépar, R. S. J. & Samset, E. (2016). Robust spatio-temporal registration of 4D cardiac ultrasound sequences. *Medical Imaging 2016: Ultrasonic Imaging and Tomography*, 9790, 97900F.

Biau, D. J., Jolles, B. M. & Porcher, R. (2010). P value and the theory of hypothesis testing: An explanation for new researchers. *Clinical Orthopaedics and Related Research®*, 468(3), 885–892.

Bishop, C., Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Carneiro, G. & Jepson, A. D. (2003). Multi-scale phase-based local features. *Computer Vision and Pattern Recognition, 2003. 2003 IEEE Computer Society Conference on*, 1, I–I.

Chaddad, A., Desrosiers, C., Toews, M. & Abdulkarim, B. (2017a). Predicting survival time of lung cancer patients using radiomic analysis. *Oncotarget*, 8(61), 104393.

Chaddad, A., Naisiri, B., Pedersoli, M., Granger, E., Desrosiers, C. & Toews, M. (2017b). Modeling Information Flow Through Deep Neural Networks. *arXiv preprint arXiv:1712.00003*.

Chaddad, A., Daniel, P., Desrosiers, C., Toews, M. & Abdulkarim, B. (2018). Novel radiomic features based on joint intensity matrices for predicting glioblastoma patient survival time. *IEEE journal of biomedical and health informatics*, 23(2), 795–804.

Chaddad, A., Toews, M., Desrosiers, C. & Niazi, T. (2019). Deep Radiomic Analysis Based on Modeling Information Flow in Convolutional Neural Networks. *IEEE Access*, 7, 97242–97252.

Chang, C.-C., Lui, C.-C., Lee, C.-C., Chen, S.-D., Chang, W.-N., Lu, C.-H., Chen, N.-C., Chang, A. Y., Chan, S. H. & Chuang, Y.-C. (2012). Clinical significance of serological biomarkers and neuropsychological performances in patients with temporal lobe epilepsy. *BMC neurology*, 12(1), 15.

Chauvin, L., Toews, M., Colliot, O., Desrosiers, C. et al. (2017). Multi-modal analysis of genetically-related subjects using SIFT descriptors in brain MRI.

Chauvin, L., Kumar, K., Desrosiers, C., De Guise, J. & Toews, M. (2018). Diffusion Orientation Histograms (DOH) for diffusion weighted image analysis. In *Computational Diffusion MRI* (pp. 91–99). Springer, Cham.

Chauvin, L., Kumar, K., Desrosiers, C., De Guise, J., Wells, W. & Toews, M. (2019). Analyzing brain morphology on the bag-of-features manifold. *International Conference on Information Processing in Medical Imaging*, pp. 45–56.

Chauvin, L., Kumar, K., Wachinger, C., Vangel, M., de Guise, J., Desrosiers, C., Wells, W., Toews, M., Initiative, A. D. N. et al. (2020). Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. *NeuroImage*, 204, 116208.

Cheung, W. & Hamarneh, G. (2009). *n*-SIFT: *n*-Dimensional Scale Invariant Feature Transform. *IEEE Transactions on Image Processing*, 18(9), 2012–2021.

Choy, C. B., Gwak, J., Savarese, S. & Chandraker, M. (2016). Universal correspondence network. *Advances in Neural Information Processing Systems*, pp. 2414–2422.

Cireşan, D., Meier, U. & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21–27.

Fergus, R., Perona, P. & Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *International journal of computer vision*, 71(3), 273–303.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1), 87–94.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics* (pp. 66–70). Springer.

Flitton, G. T., Breckon, T. P. & Bouallagu, N. M. (2010). Object Recognition using 3D SIFT in Complex CT Volumes. *BMVC*, (1), 1–12.

Friendly, M. & Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. *URL http://www. datavis. ca/milestones*, 32, 13.

Frisken, S., Luo, M., Machado, I., Unadkat, P., Juvekar, P., Bunevicius, A., Toews, M., Wells, W., Miga, M. I. & Golby, A. J. (2019a). Preliminary results comparing thin-plate splines with finite element methods for modeling brain deformation during neurosurgery using intraoperative ultrasound. *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, 10951, 1095120.

Frisken, S., Luo, M., Juvekar, P., Bunevicius, A., Machado, I., Unadkat, P., Bertotti, M. M., Toews, M., Wells, W. M., Miga, M. I. et al. (2019b). A comparison of thin-plate spline deformation and finite element modeling to compensate for brain shift during tumor resection. *International journal of computer assisted radiology and surgery*, 1–11.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2), 119–130.

Gan, E. & Bailis, P. (2017). Scalable kernel density classification via threshold-based pruning. *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 945–959.

Gill, G., Toews, M. & Beichel, R. R. (2014). Robust initialization of active shape models for lung segmentation in CT scans: a feature-based atlas approach. *International journal of biomedical imaging*, 2014.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Grimson, W. E. L. & Lozano-Perez, T. (1987). Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4), 469–482.

Haacke, E. M., Brown, R. W., Thompson, M. R., Venkatesan, R. et al. (1999). *Magnetic resonance imaging: physical principles and sequence design*. Wiley-Liss New York:.

Haralick, R. M., Shanmugam, K. & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610–621.

Harris, C. & Stephens, M. (1988). A combined corner and edge detector. *Alvey vision conference*, 15(50), 10–5244.

Hartley, R. & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

Jian, B. & Vemuri, B. C. (2011). Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8), 1633–1645.

Kadir, T. & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, 45(2), 83–105.

Ke, Y. & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2, II–II.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J. & Ziegler, H. (2008). Visual analytics: Scope and challenges. In *Visual data mining* (pp. 76–90). Springer.

Keller, P. R., Keller, M. M., Markel, S., Mallinckrodt, A. J. & McKay, S. (1994). Visual cues: practical data visualization. *Computers in Physics*, 8(3), 297–298.

Kim, D., Kim, K., Kim, J.-Y., Lee, S., Lee, S.-J. & Yoo, H.-J. (2009). 81.6 GOPS object recognition processor based on a memory-centric NoC. *IEEE transactions on very large scale integration (VLSI) systems*, 17(3), 370–383.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*,

pp. 1097–1105.

Krzyzak, A. & Pawlak, M. (1984). Distribution-free consistency of a nonparametric kernel regression estimate and classification. *IEEE Transactions on Information Theory*, 30(1), 78–81.

Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O. & Toews, M. Fiberprint: Identifying subjects and twins using fiber geometry based brain fingerprint.

Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O. & Toews, M. (2017a). L'empreinte cérébrale: une image des connexions du cerveau. *Substance ÉTS*.

Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O. & Toews, M. (2017b). Fiberprint: Human Brain Wiring Shows Unique Fingerprint. *Substance ÉTS*.

Kumar, K., Toews, M., Chauvin, L., Colliot, O. & Desrosiers, C. (2018). Multi-modal brain fingerprinting: a manifold approximation based framework. *NeuroImage*, 183, 212–226.

Laguna, A. B., Riba, E., Ponsa, D. & Mikolajczyk, K. (2019). Key. Net: Keypoint Detection by Handcrafted and Learned CNN Filters. *arXiv preprint arXiv:1904.00889*.

Langenfeld, F., Axenopoulos, A., Chatzitofis, A., Craciun, D., Daras, P., Du, B., Giachetti, A., Lai, Y.-k., Li, H., Li, Y. et al. (2018). SHREC 2018–Protein Shape Retrieval. *Eurographics Workshop on 3D Object Retrieval*, pp. 53–61.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989a). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.

LeCun, Y. et al. (1989b). Generalization and network design strategies. *Connectionism in perspective*, 143–155.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International journal of computer vision*, 30(2), 79–116.

Lindeberg, T. (2013). *Scale-space theory in computer vision*. Springer Science & Business Media.

Lindeberg, T., Lidberg, P. & Roland, P. (1999). Analysis of brain activation patterns using a 3-D scale-space primal sketch. *Human brain mapping*, 7(3), 166–94.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., Toews, M., Unadkat, P., Sedghi, A., Zhou, H. et al. (2018a). Using the variogram for vector outlier screening:

application to feature-based image registration. *International journal of computer assisted radiology and surgery*, 13(12), 1871–1880.

Luo, J., Frisken, S., Popuri, K., Cobzas, D., Preiswerk, F., Toews, M., Zhang, M., Ding, H., Golland, P., Golby, A. et al. (2018b). On the Ambiguity of Registration Uncertainty. *arXiv preprint arXiv:1803.05266*.

Luo, J., Toews, M., Machado, I., Frisken, S., Zhang, M., Preiswerk, F., Sedghi, A., Ding, H., Pieper, S., Golland, P. et al. (2018c). A feature-driven active framework for ultrasound-based brain shift compensation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 30–38.

Lupovitch, J. (2006). In the Blink of an Eye: How Vision Sparked the Big Bang of Evolution. *Archives of Ophthalmology - ARCH OPHTHALMOL*, 124, 142-142. doi: 10.1001/archopht.124.1.142.

Machado, I., Toews, M., Luo, J., Unadkat, P., Essayed, W., George, E., Teodoro, P., Carvalho, H., Martins, J., Golland, P. et al. (2018a). Deformable mri-ultrasound registration via attribute matching and mutual-saliency weighting for image-guided neurosurgery. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation* (pp. 165–171). Springer, Cham.

Machado, I., Toews, M., Luo, J., Unadkat, P., Essayed, W., George, E., Teodoro, P., Carvalho, H., Martins, J., Golland, P. et al. (2018b). Non-rigid registration of 3D ultrasound for neurosurgery using automatic feature detection and matching. *International journal of computer assisted radiology and surgery*, 13(10), 1525–1538.

Machado, I., Toews, M., George, E., Unadkat, P., Essayed, W., Luo, J., Teodoro, P., Carvalho, H., Martins, J., Golland, P. et al. (2019). Deformable MRI-Ultrasound registration using correlation-based attribute matching for brain shift correction: Accuracy and generality in multi-site data. *NeuroImage*, 116094.

Mahalanobis, P. C. (1936). On the generalized distance in statistics.

Maintz, J. A. & Viergever, M. A. (1998). A survey of medical image registration. *Medical image analysis*, 2(1), 1–36.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C. & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9), 1498–1507.

Marr, D. & Poggio, T. (1977). *A Theory of Human Stereo Vision.*

Masoumi, M., Lombaert, H. & Toews, M. (2019). WaveletBrain: Characterization of human brain via spectral graph wavelets. *arXiv preprint arXiv:1906.06158*.

McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K. & Trus, B. L. (2001). Medical image processing, analysis and visualization in clinical research. *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, pp. 381–386.

Mei, K., Dong, P., Lei, H. & Fan, J. (2014). A distributed approach for large-scale classifier training and image classification. *Neurocomputing*, 144, 304–317.

Meng, X.-L. et al. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3), 1142–1160.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. & Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pp. 41–48.

Mikolajczyk, K. & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1), 63–86.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. & Van Gool, L. (2005). A comparison of affine region detectors. *International journal of computer vision*, 65(1-2), 43–72.

Moravec, H. P. (1979). Visual mapping by a robot rover. *Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1*, pp. 598–600.

Muja, M. & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340), 2.

Neyman, J. & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706), 289–337.

Ng, B., Toews, M., Durrleman, S., Shi, Y., Gao, F., Shi, P., Farag, A. A., Shalaby, A., El Munim, H. A., Farag, A. et al. Part I Methods and Models.

Nguyen, A., Yosinski, J. & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436.

Olejnik, S., Li, J., Supattathum, S. & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of educational and behavioral statistics*, 22(4), 389–406.

Ono, Y., Trulls, E., Fua, P. & Yi, K. M. (2018). LF-Net: learning local features from images. *Advances in Neural Information Processing Systems*, pp. 6234–6244.

Pennec, X., Ayache, N. & Thirion, J.-P. (2000). Landmark-based registration using features identified through differential geometry. Academic Press.

Piringer, H. (2011). *Large data scalability in interactive visual analysis*. (Ph.D. thesis, Piringer).

Póczos, B., Xiong, L., Sutherland, D. J. & Schneider, J. (2012). Nonparametric kernel estimators for image classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2989–2996.

Preacher, K. & Briggs, N. (2001). Calculation for Fisher's Exact Test: An interactive calculation tool for Fisher's exact probability test for 2× 2 tables. *Chapel Hill NC: University of North Carolina*.

Rohr, K. (1997). On 3D differential operators for detecting point landmarks. *Image and Vision Computing*, 15(3), 219–233.

Rohr, K., Stiehl, H. S., Sprengel, R., Buzug, T. M., Weese, J. & Kuhn, M. (2001). Landmark-based elastic registration using approximating thin-plate splines. *IEEE Transactions on medical imaging*, 20(6), 526–534.

Romeny, B. M. H. (2008). *Front-end vision and multi-scale image analysis: multi-scale computer vision theory and applications, written in mathematica*. Springer Science & Business Media.

Salzwedel, A. P., Grewen, K. M., Vachet, C., Gerig, G., Lin, W. & Gao, W. (2015). Prenatal drug exposure affects neonatal brain functional connectivity. *Journal of Neuroscience*, 35(14), 5860–5869.

Scovanner, P., Ali, S. & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360.

Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3), 683–690.

Shen, W., Neyman, J., Pearson, E., Bolch, G., Greiner, S., de Meer, H., Trivedi, K., Sahner, R., Trivedi, K., Puliafito, A. et al. (2018). On the problem of the most efficient tests of

statistical hypotheses. *Interfaces*, 48(3), 1–5.

Shi, J. et al. (1994). Good features to track. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pp. 593–600.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Simon, I., Snavely, N. & Seitz, S. M. (2007). Scene summarization for online image collections. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8.

Sinha, P. & Poggio, T. (2002). 'United'we stand. *Perception*, 31(1), 133.

Skurichina, M. & Duin, R. P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2), 121–135.

Stewart, C. V., Tsai, C.-L. & Roysam, B. (2003). The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE transactions on medical imaging*, 22(11), 1379–1394.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.

Su, Y., Shan, S., Chen, X. & Gao, W. (2009). Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Transactions on image processing*, 18(8), 1885–1896.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Talairach, J. & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging.

Terrell, G. R., Scott, D. W. et al. (1992). Variable kernel density estimation. *The Annals of Statistics*, 20(3), 1236–1265.

Toews, M. (2002). Entropy-of-likelihood feature point selection for image correspondence.

Toews, M. & Arbel, T. (2007). A statistical parts-based model of anatomical variability. *IEEE Transactions on Medical Imaging*, 26(4), 497–508.

Toews, M. & Arbel, T. (2009). Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9), 1567–1581.

Toews, M. & Wells, W. (2009). SIFT-Rank: Ordinal description for invariant feature correspondence. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 172–177.

Toews, M. & Wells, W. M. (2010). A mutual-information scale-space for image feature detection and feature-based classification of volumetric brain images. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 111–116.

Toews, M. & Wells, W. M. (2013). Efficient and robust model-to-image alignment using 3D scale-invariant features. *Medical image analysis*, 17(3), 271–282.

Toews, M. & Wells, W. M. (2016). How are siblings similar? how similar are siblings? large-scale imaging genetics using local image features. *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pp. 847–850.

Toews, M. & Wells, W. M. (2017). Phantomless auto-calibration and online calibration assessment for a tracked freehand 2-D ultrasound probe. *IEEE transactions on medical imaging*, 37(1), 262–272.

Toews, M., Wells, W., Collins, D. L. & Arbel, T. (2010). Feature-based morphometry: Discovering group-related anatomical patterns. *NeuroImage*, 49(3), 2318–2327.

Toews, M., Wells, W. M. & Zöllei, L. (2012). A feature-based developmental model of the infant brain in structural MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 204–211.

Toews, M., Zöllei, L. & Wells, W. M. (2013). Feature-based alignment of volumetric multi-modal images. *International Conference on Information Processing in Medical Imaging*, pp. 25–36.

Toews, M., Wachinger, C., Estepar, R. S. J. & Wells, W. M. (2015). A feature-based approach to big data analysis of medical images. *International Conference on Information Processing in Medical Imaging*, pp. 339–350.

Tran, T. N., Wehrens, R. & Buydens, L. M. (2006). KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis*, 51(2), 513–525.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.

Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 395–402.

Urschler, M., Zach, C., Ditt, H. & Bischof, H. (2006). Automatic point landmark matching for regularizing nonlinear intensity registration: Application to thoracic CT images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 710–717.

Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhyā: The Indian Journal of Statistics, Series A*, 261–270.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1, I–I.

Wachinger, C., Toews, M., Langs, G., Wells, W. & Golland, P. (2015). Keypoint transfer segmentation. *International Conference on Information Processing in Medical Imaging*, pp. 233–245.

Wachinger, C., Toews, M., Langs, G., Wells, W. & Golland, P. (2018). Keypoint transfer for fast whole-body segmentation. *IEEE transactions on medical imaging*.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. Taylor & Francis.

Wells III, W. M. (1997). Statistical approaches to feature-based object recognition. *International Journal of Computer Vision*, 21(1-2), 63–98.

Wikipedia, U. . W. . U.-P. L. . [File:P value.png]. (2014). Illustration of describing the meaning of p-value in statistical significance testing. Consulted at https://commons.wikimedia.org/wiki/File:P-value_in_statistical_significance_testing.svg.

Wolverton, C. & Wagner, T. (1969). Asymptotically optimal discriminant functions for pattern classification. *IEEE Transactions on Information Theory*, 15(2), 258–265.

Yang, J., Williams, J. P., Sun, Y., Blum, R. S. & Xu, C. (2011). A robust hybrid method for nonrigid image registration. *Pattern Recognition*, 44(4), 764–776.

Yi, K. M., Trulls, E., Lepetit, V. & Fua, P. (2016). Lift: Learned invariant feature transform. *European Conference on Computer Vision*, pp. 467–483.

Zhang, Z., Xie, Y., Xing, F., McGough, M. & Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. *International Conference on Computer Vision and Pattern Recognition, arXiv preprint*.

Zheng, L., Yang, Y. & Tian, Q. (2017). SIFT meets CNN: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5), 1224–1244.