



**I
N
A
O
E**

**Feature extraction and classification of a two-class fMRI
experiment using principal component analysis, LDA
and SVM**

Juan Manuel Ramírez Cortés
Electronics Department
INAOE
Tonantzintla, Puebla
Mexico

Jingqi Ao
Computer Vision and Image Analysis Laboratory
Electrical and Computing Engineering Department
Texas Tech University
Lubbock, USA

Reporte Técnico No. RT-0590
RT
R36
2009

Coordinación de Electrónica
INAOE
Calle Luis Enrique Erro No. 1
Sta. Ma. Tonantzintla
Cholula, Puebla, 72840
México

©INAOE 2009

Derechos Reservados

El autor otorga al INAOE el permiso de reproducir y
distribuir copias de este reporte en su totalidad o en partes.



This report is the result of the research developed during the summer research stay of the first author at the Computer Vision and Image Analysis Laboratory, Electrical and Computing Engineering Department, Texas Tech University, USA. The first author worked with the graduate student Jingqi Ao who obtained the results described in this report as part of his graduate studies.

ACKNOWLEDGEMENTS:

fMRI data provided by Dr. Michael O'Boyle.
Research developed in the Computer Vision and Image Analysis Laboratory.
Directors: Dr. Brian Nutter; Dr. Sunanda Mitra.

August 20th, 2009

INDEX

	Page
Abstract	4
1. Introduction	5
1.1 Magnetic Resonance Imaging (MRI).	5
1.2 Functional Magnetic Resonance Imaging.	6
1.3 FSL (FMRIB's Software Library)	8
1.4 FEAT; Model-based FMRI analysis	
2. Feature extraction and classification techniques	9
2.1 Feature extraction and feature selection	10
2.2 Classification techniques	11
2.3 GLM-based t-test feature extraction method	12
2.4 Two-sample t-test feature extraction method	15
3. Experimental work and results.	17
3.1 Preprocessing.	17
3.2 Results of both feature extraction methods.	18
3.3 Block diagram of the whole algorithm	19
3.4 Reorganization of Data in PCA Input Form	21
3.5 Feature Extraction – PCA	21
3.6 Feature Selection-Max/ForwardSearch	21
3.7 Training and Testing the Classifier	22
3.8 Feature Extraction – GLM-based t-test/Two-sample t-test	24
4. Results	25
4.1 Extracted features from different feature extraction methods	25
4.2 Selected features from different feature selection methods	27
4.3 Accuracy rate from different feature extraction and selection methods	30
4.4 Results on classification	31
Conclusion	32
Acknowledgments	33
References	34
Appendix A Principal Component Analysis	36
Appendix B Support Vector Machine	37
Appendix C Linear Discriminant Analysis	38
Appendix D K-Nearest Neighbor	39

Abstract

Functional magnetic resonance imaging (fMRI), is a technique for investigating brain's activity in a mental process, in response to some specific applied stimulus. In this report, a two-class fMRI classification project is described. fMRI (Functional Magnetic Resonance Imaging) image is inherently high-dimensional. More than ten millions voxels in the raw fMRI data make the fMRI-related pattern recognition difficult to perform directly. Different methods have been proposed to solve this problem, such as raw-data related t-value¹ method, Principal Component Analysis (PCA)² and Independent Component Analysis (ICA)³. In this project, PCA (Principal Component Analysis) is used as feature extraction method to decrease the dimension of fMRI data. Two feature selection methods are used: Max-group and Forward-Search. Three most commonly used classifiers in two-class classification are compared: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (KNN). Results obtained from experiments using preprocessed (high-pass filtering and motion-correction) and non-preprocessed data from 8 different subjects, are presented. Finally, two more feature extraction methods (GLM-based t-test and Two-sample t-test) are proposed to improve the accuracy rate of classification. Feature selection methods are kept unchanged, still containing MaxGroup and ForwardSearch. The features from different feature extraction methods and different feature selection methods are compared in the form of taking one subject's result as an example. Accuracy Rate of 8 different subjects, obtained with 4 combinations of feature extraction methods and feature selection methods, are finally presented.

1. Introduction

1.1 Magnetic Resonance Imaging (MRI).

A magnetic resonance imaging scan of the head is a noninvasive method to create detailed pictures of the brain and surrounding nerve tissues. MRI Technology was introduced in 1977 as a medical imaging technique used in radiology to visualize detailed internal structure of the body. The body is mainly composed of water molecules which each contain two hydrogen nuclei or protons. When a person goes inside the powerful magnetic field of an MRI scanner, these protons align with the direction of the field. A second radio frequency electromagnetic field is then briefly turned on causing the protons to absorb some of its energy. When this field is turned off the protons release this energy at a radio frequency which can be detected by the scanner. The position of protons in the body can be determined by applying additional magnetic fields during the scan which allows an image of the body to be built up. These are created by turning gradients coils on and off which creates the knocking sounds heard during an MR scan [1,2]. Diseased tissue, such as tumors, can be detected because the protons in different tissues return to their equilibrium state at different rates. By changing the parameters on the scanner this effect is used to create contrast between different types of body tissue. Figure 1 shows a typical MRI brain image.

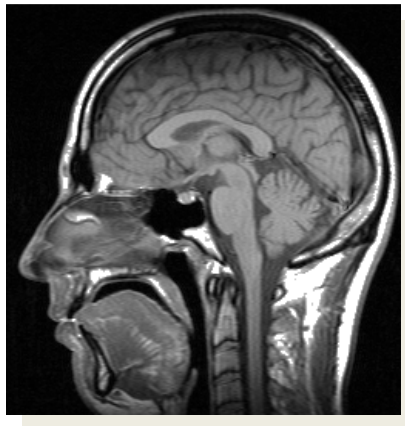


Figure 1 MRI brain image

1.2 Functional Magnetic Resonance Imaging.

Functional MRI (fMRI) is based on BOLD (Blood Oxygen-Level Dependent) phenomenon in the brain, which can provide some indication of brain activities [3]. Generally, fMRI image is processed by HRF (Haemodynamic Response Function)-based statistical analysis, using software such as FSL and SPM [4]. Functional MRI (fMRI) measures signal changes in the brain that are due to changing neural activity. The brain is scanned at low resolution but at a rapid rate (typically once every 2–3 seconds). The object of this analysis is to find the activation areas in the brain, which correspond to some stimulus in a time axis. There are also other methods to utilize fMRI image, for example, using the computer to classify the fMRI images from different stimuli, namely “Cognition State Detection” [5]. The work in this report is to perform classification of fMRI images obtained from a two-states experiment focusing on feature extraction and selection methods. The experiment is based on the fMRI scanning of some subjects exposed to two different stimuli: counting and reasoning. The two stimuli are applied consecutively in a previously defined order, which conforms a group of 144 three-dimensional volumes. Figure 2 shows a representation of a typical model-based fMRI analysis. Each voxel in the three-dimensional representation of the brain generates a time series when is analyzed across the time axis. This time series is correlated with the theoretical model in order to detect the activation voxels corresponding to the specific stimulus, as represented in Figure 3. Figure 4 corresponds to one slice of some typical fMRI image with the activation voxels represented in pseudocolor.

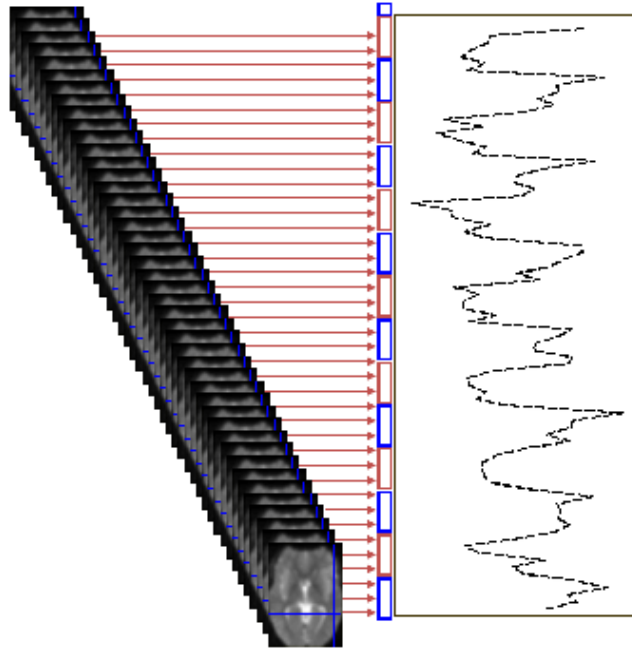


Figure 2. Representation of a typical model-based fMRI analysis

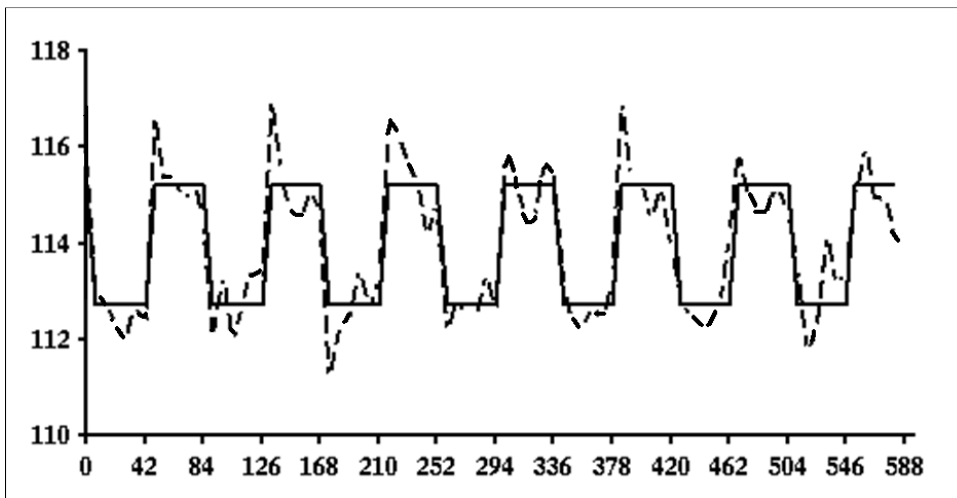


Figure 3. Theoretical model superimposed to a time series obtained from a fMRI experiment.

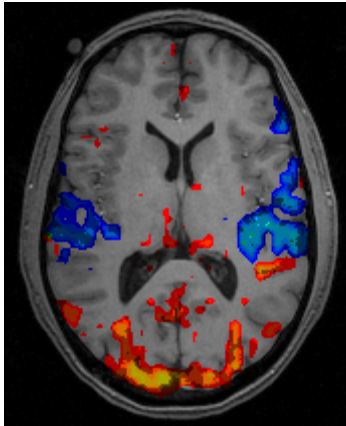


Figure 4. Typical fMRI image showing the activation voxels of a two-class experiment

1.3 FSL (FMRIB's Software Library)

FSL is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. FSL was developed by the Analysis Group at the Oxford Centre for Functional MRI of the Brain (FMRIB) [6]. FSL has been made available to the wider community as a single integrated software package, FMRIB's Software Library (FSL). FSL is available as both source code and as self-contained binary distributions for Linux, MacOS X (Apple), Windows XP (under Cygwin), Solaris (Sun), and IRIX (Silicon Graphics). It is freely available for academic (noncommercial) use. Almost all tools can be run both from the command line and via GUIs. The FSL binary distributions come with all necessary third-party software bundled, making installation and setup extremely easy. FSL is in active use in over 1000 laboratories worldwide. Most FSL tools come with web-based documentation, and the most complex tools, such as FEAT, include comprehensive bubble-popup help in the GUI. Example data are supplied in a separate download named FSL Evaluation and Example Data Suite (FEEDS), which as well as providing example data for trying out the various FSL tools includes a script that runs the core tools automatically and tests the results against supplied results to check that FSL is running correctly. Further support on using FSL is offered via the FSL email list at www.jiscmail.ac.uk/lists/fsl.html. FSL was used in this work for the pre-processing part: Motion correction, filtering and registration to the standard model.

1.4 FEAT; Model-based FMRI analysis

FEAT is a software tool for high quality model-based FMRI data analysis, with an easy-to-use graphical user interface [6]. FEAT is part of FSL (FMRIB's Software Library). FEAT automates as many of the analysis decisions as possible, and allows easy (though still robust, efficient and valid) analysis of simple experiments whilst giving enough flexibility to also allow sophisticated analysis of the most complex experiments. The data modelling which FEAT uses is based on general linear modelling (GLM), otherwise known as multiple regression. It allows you to describe the experimental design; then a model is created that should fit the data, telling you where the brain has activated in response to the stimuli. In FEAT, the GLM method used on first-level (time-series) data is known as FILM (FMRIB's Improved Linear Model). FILM uses a robust and accurate nonparametric estimation of time series autocorrelation to prewhiten each voxel's time series; this gives improved estimation efficiency compared with methods that do not pre-whiten. General Linear Modelling (more correctly known simply as "linear modelling") sets up a model (i.e., what you expect to see in the data) and fits it to the data. If the model is derived from the stimulation that was applied to the subject in the MRI scanner, then a good fit between the model and the data means that the data was indeed caused by the stimulation. When the model is fit to the data, for each voxel there will be found an estimate of the "goodness of fit" of each column in the model, to that voxel's timecourse.

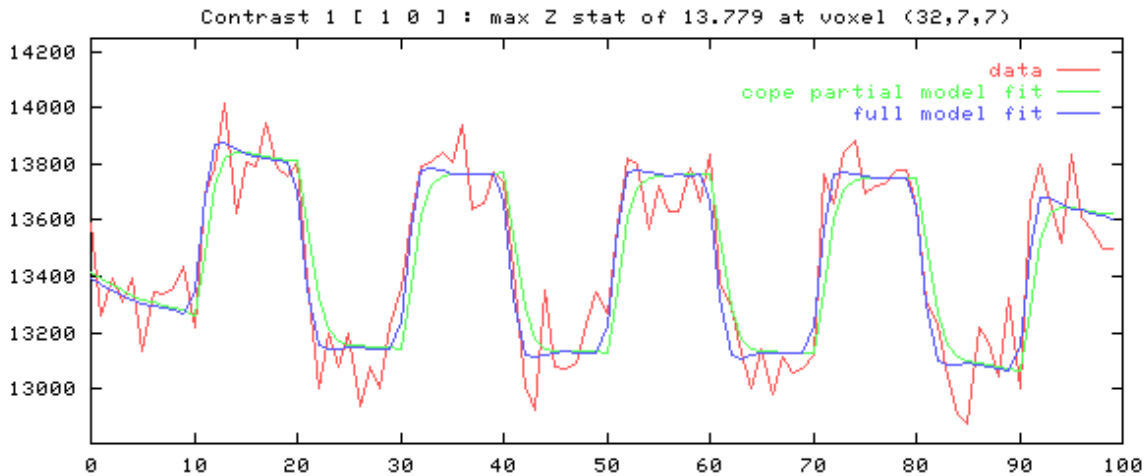


Figure 5. An example of a voxel time course obtained from FSL

2. Feature extraction and classification techniques

2.1 Feature extraction and feature selection

The goal of feature extraction is to decrease fMRI data's dimension. The feature extraction is quite important for fMRI image classification, because fMRI is inherent high-dimensional. Typically, one volume of fMRI contains about 150,000 voxels, and one experiment will include more than 100 volumes. It means that the raw dataset of one fMRI experiment will contain more than ten million voxels. Processing all these voxels as features for classification is practically impossible because of memory and time limitation. It is necessary to do some transformation or use only parts of these voxels to decrease our fMRI data's dimension. Feature selection is performed after the feature extraction. The object of feature selection is to use parts of extracted features as input of classifier. In many cases, even after the feature extraction, the number of features is still too large to feed the classifier. Feature selection is needed to perform to filter our extracted features and find the group of features which can get the best classification performance. In this report, pattern recognition of fMRI is performed with PCA-based feature extraction. The key point of choosing PCA as feature extraction method is its speed. Many controversies about PCA

implemented in pattern recognition rise from its classification-blinded characteristic⁴. It will be demonstrated that the feature selection method which follows the PCA-based feature extraction can help to improve the performance of PCA-based classification. Two different feature selection methods are used: Max-group algorithm and Forward-search (which is also called “climbing hill”) algorithm.

2.2 Classification techniques

Three kinds of classifiers which are most commonly used in two-class classification are compared. Support Vector Machine (SVM) is the newly-emerging classifier and get very good performance in many cases. Linear Discriminant Analysis (LDA) is the traditional method used in two-class classification. And k-Nearest Neighbor has the simplest algorithm and in many cases it works well. Sometimes, the preprocessing of image will affect the performance of classification. Two kinds of fMRI functional images are used: Unfiltered-fMRI function images (that is, the raw data) and Filtered-fMRI function images (that is, images processed by FSL to eliminate the motion-effect and to attenuate low-frequency noise). As the original dimension of fMRI data-space is very high (nearly 150,000 voxels in one volume), before doing the classification, feature extraction and selection is required. Feature extraction is to decrease fMRI data’s dimension. Two main categories of feature extraction of fMRI are transformation-based and voxel-based. Transformation-based feature extraction uses some transformation to decrease the dimension of fMRI data. Generally, these transformations project the data from original, high-dimensional coordinate system to a new, low-dimensional coordinate system. Transformation-based feature extraction includes PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). Voxel-based feature extraction keeps the data in the original coordinate system. However, it will rank all the voxels and take only high-rank voxels as the result of the feature extraction. For example, even the original data dimension is 15000 (15000 voxels in one 3D volume), voxel-based feature extraction will grade all 15000 voxels and only pick the top 1000 voxels as the extracted features, so real data dimension decreases from 15000 to 1000. Feature selection is performed after the feature

⁴ Pattern Recognition, Robi Polikar, Wiley Encyclopedia of Biomedical Engineering, 2006

extraction. The object of feature selection is to use parts of extracted features as input of classifier. More details about feature selection can be obtained in the report “Pattern recognition of fMRI image using PCA method.”

In this report, pattern recognition of fMRI is performed with two different feature extraction methods: GLM (General Linear Model)-based t-test and Two-sample t-test. They can be categorized as voxel-based feature extraction.

2.3 GLM-based t-test feature extraction method

GLM (General Linear Model)-based feature extraction method uses the GLM-based statistical analysis’s result. Most of the fMRI analysis tools, such as FSL and SPM, use GLM and Contrast of Parameter Estimate (COPE) to estimate the relationship between response and different stimuli in one voxel scale.

GLM describes response in one voxel as

$$Y_{n \times 1}(v) = X_{n \times p}(v)\beta_{p \times 1}(v) + \epsilon_{n \times 1}(v)$$

Where Y is a two dimensional matrix ($n \times 1$) describing the fMRI intensity in voxel v during the whole time course. Generally, n equals the number of volumes in one experiment. X is called “design matrix” and it is constituted by p different stimulus, which have the same time-length with Y (n volumes). β is called “(regression) parameter estimates” (PE). It contain p elements and each element β_i is the estimated regression parameter for $X_{n \times i}$. ϵ is called “residual error” and each element ϵ_i in ϵ corresponds to $Y_{i \times 1}(v) - X_{i \times p}(v)\beta_{p \times 1}(v)$.

Figure 1 is an example of design matrix used in analysis tool FSL. There are four columns in that design matrix. However, there are only two stimuli, counting-related stimulus (first column, marked with “C”) and reasoning-related stimulus (third column, marked with “R”). The second and fourth columns are first-order derivation of counting-related stimulus (first

column) and reasoning-related stimulus (third column). These derivations are used to improve the precision of HRF-model used in estimation.

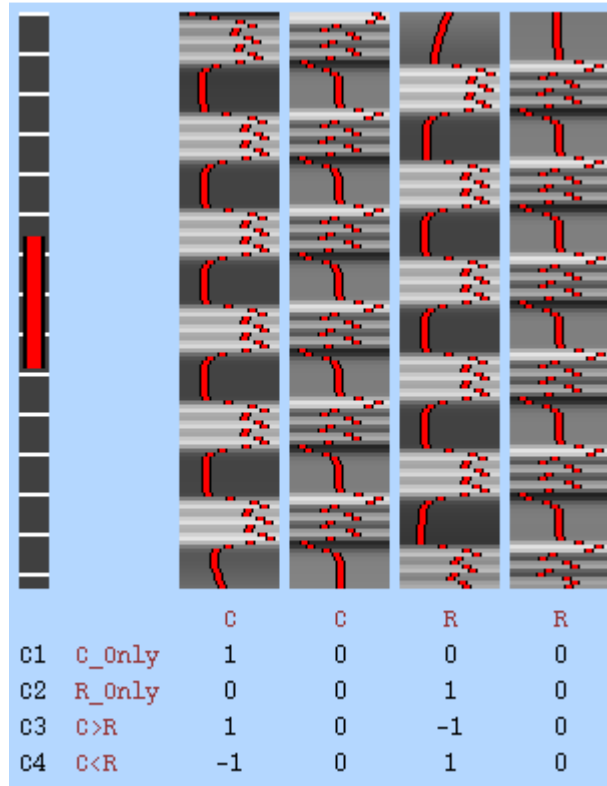


Figure 6 Example of design matrix and contrast in FSL

For most cases, there is no need to know the relationships between response and all stimuli. So Contrast of Parameter Estimate (COPE) is used to replace β . In fact, Contrast of Parameter Estimate (COPE) is a “combination of some of elements in β ”. To get a COPE, a contrast is used to “mask” the stimuli which are not interested in. For example, there are totally 3 different stimuli $X_{n \times 3}$ and corresponding $\beta_1 \sim \beta_3$, however, only β_1 and β_2 need to be estimated. So a contrast $[1 \ 1 \ 0]$ is used to mask the β_3 and maintain β_1 and β_2 via $\text{COPE} = \text{contrast} \times [\beta_i]$ (that is, $\text{COPE} = [1 \ 1 \ 0] \times [\beta_1 \ \beta_2 \ \beta_3]$). Correspondingly, $X_{n \times 3}$ will be masked too.

As Doug Greve (Greve) has pointed out, with GLM, we can get an estimation of β by:

$$\hat{\beta}(v) = (X^T X)^{-1} X^T Y, \hat{\beta} \text{ is the estimation of } \beta$$

and its variance is

$$\text{Var}(\hat{\beta}(v)) = [(X^T X)^{-1} X^T] \text{Var}(Y) [(X^T X)^{-1} X^T]^T$$

In most fMRI cases, $\hat{\beta}$ usually means estimation of COPE, $\widehat{\text{COPE}}$.

Student t-test is used to check the reliability of $\widehat{\text{COPE}}$. T-test is performed with

$$t_{\text{cope}}(v) = \frac{\widehat{\text{COPE}}(v)}{\sigma_{\widehat{\text{COPE}}(v)}}, \sigma \text{ means standard deviation}$$

The higher $t_{\text{cope}}(v)$ is, the more reliable relationship between responses with certain stimulus (stimuli) combination (represented by a specific COPE) there is in voxel v . GLM-based t-test feature extraction method will choose voxels with the high $t_{\text{cope}}(v)$ as extracted features.

It needs to point out that, for each contrast there is a specific COPE, which means one contrast corresponds to one group GLM-based t-value. As the GLM is implemented in the voxel-scale, one contrast will produce a one-volume-size GLM-based t-value matrix. For example, if our data is constituted n volumes and the size of each volume is $64*64*35$. One contrast, like $[1 \ 1 \ 0]$ used above, will produce a GLM-based t-value volume, which size is also $64*64*35$. It means that if we use k contrasts, we will have k volumes of t-values.

2.4 Two-sample t-test feature extraction method

Two-sample t-test feature extraction is based on the work of Serdar K. Balci (Balci, 2008), and Didier G. Leibovici (D.G. Leibovici, 2001) et al. This method is based on two-sample t-test.

In our case, for each voxel, the two-sample t-test is performed by calculating $t(v)$:

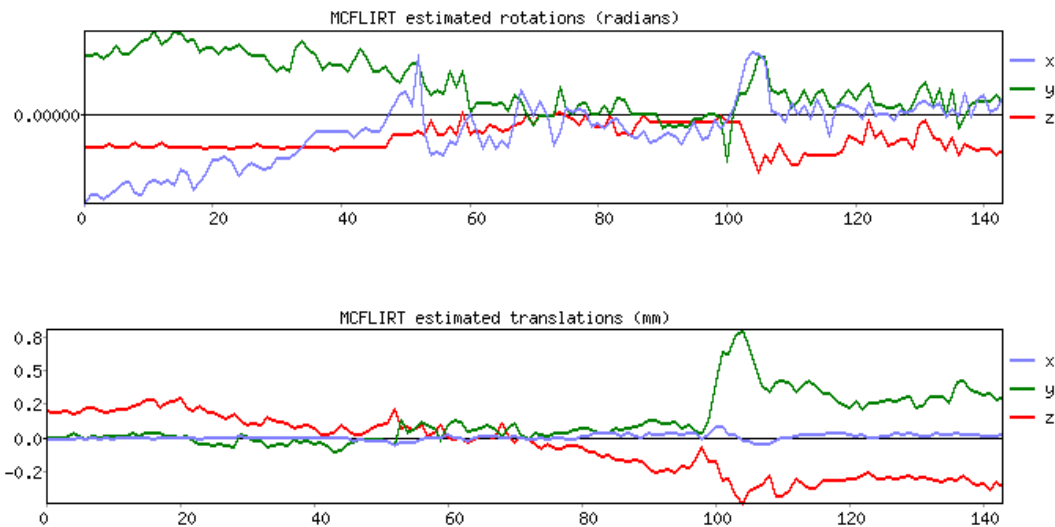
$$t(v) = \frac{\mu_{\text{counting}}(v) - \mu_{\text{reasoning}}(v)}{\sqrt{\frac{\sigma_{\text{counting}}^2(v)}{n_{\text{counting}}} + \frac{\sigma_{\text{reasoning}}^2(v)}{n_{\text{reasoning}}}}} \quad (\text{Eq.1})$$

μ is the mean of intensity of fMRI in voxel v through all counting/reasoning related volumes. σ is the standard deviation of intensity of fMRI in voxel v through all counting/reasoning related volumes. n is the number of volumes related with counting/reasoning. Generally speaking, the larger $t(v)$ is, the larger difference between the responses for different stimuli exists in this voxel v . As that difference will be critical in two-class classification, two-sample t-test feature extraction will take the voxel whose $t(v)$ is large as extracted features.

3. Experimental work and results.

3.1 Preprocessing.

FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 5.98, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Z (Gaussianised T/F) statistic images were thresholded at $P=0.05$ (uncorrected). The following pre-statistics processing was applied; motion correction using MCFLIRT non-brain removal using BET, spatial smoothing using a Gaussian kernel of FWHM 8.0mm; grand-mean intensity normalisation of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma=36.0s$). MCFLIRT Motion correction and registration. Mean displacements: absolute=0.29mm, relative=0.06mm Summary registration, FMRI to standard space registration to high resolution structural and/or standard space images was carried out using FLIRT.



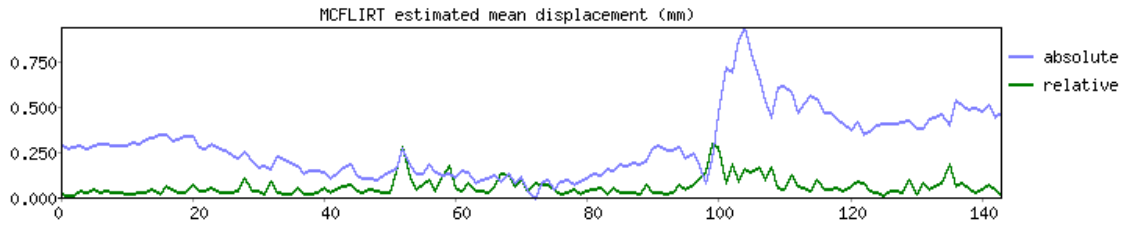


Figure 7. Estimated rotation, translation, and displacement from motion correction

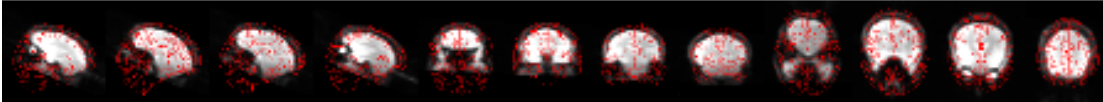


Figure 8. Registration to standar space.

3.2 Results of both feature extraction methods.

As described above, in both feature extraction methods (GLM-based t-test and Two-sample t-test), the voxels with “high $t_{cope}(v)$ ” or “large $t(v)$ ” are the subjects to be chosen as extracted features. However, there is no common rule to define how much is “high” or “large” enough to be chosen. In this report, considering the limitation of memory and time, both feature extraction method will choose voxels with top 1000 t-values as extracted features.

In this report, two feature selection methods are the same as the methods used in the last report “Pattern recognition of fMRI image using PCA method”. Max-group feature selection method just chooses top voxels in extracted features (top 10 voxels in this report). Forward-search feature selection method searches all the extracted features and finds the best performance features one by one to form the final 10-feature group. There is a more detailed description in the last report “Pattern recognition of fMRI image using PCA method”. Support Vector Machine (SVM) is used as classifier in pattern recognition. There is a detailed appendix (appendix B) in the last report “Pattern recognition of fMRI image using PCA method”. fMRI images are images preprocessed by FSL. The preprocessing includes motion-correct realignment and highpass filtering. The performance will be demonstrated with 8 different subjects.

3.3 Block diagram of the whole algorithm

As the original dimension of fMRI data-space is very high (nearly 150,000 voxels in one volume), before doing the classification, feature extraction and selection is required. Feature extraction is to decrease fMRI data's dimension. Two main categories of feature extraction of fMRI are transformation-based and voxel-based. Transformation-based feature extraction uses some transformation to decrease the dimension of fMRI data. Generally, these transformations project the data from original, high-dimensional coordinate system to a new, low-dimensional coordinate system. Transformation-based feature extraction includes PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). Voxel-based feature extraction keeps the data in the original coordinate system. However, it will rank all the voxels and take only high-rank voxels as the result of the feature extraction. For example, even the original data dimension is 15000 (15000 voxels in one 3D volume), voxel-based feature extraction will grade all 15000 voxels and only pick the top 1000 voxels as the extracted features, so real data dimension decreases from 15000 to 1000. Feature selection is performed after the feature extraction. The object of feature selection is to use parts of extracted features as input of classifier. More details about feature selection can be obtained in the report "Pattern recognition of fMRI image using PCA method." In this report, pattern recognition of fMRI is performed with two different feature extraction methods: GLM (General Linear Model)-based t-test and Two-sample t-test. They can be categorized as voxel-based feature extraction.

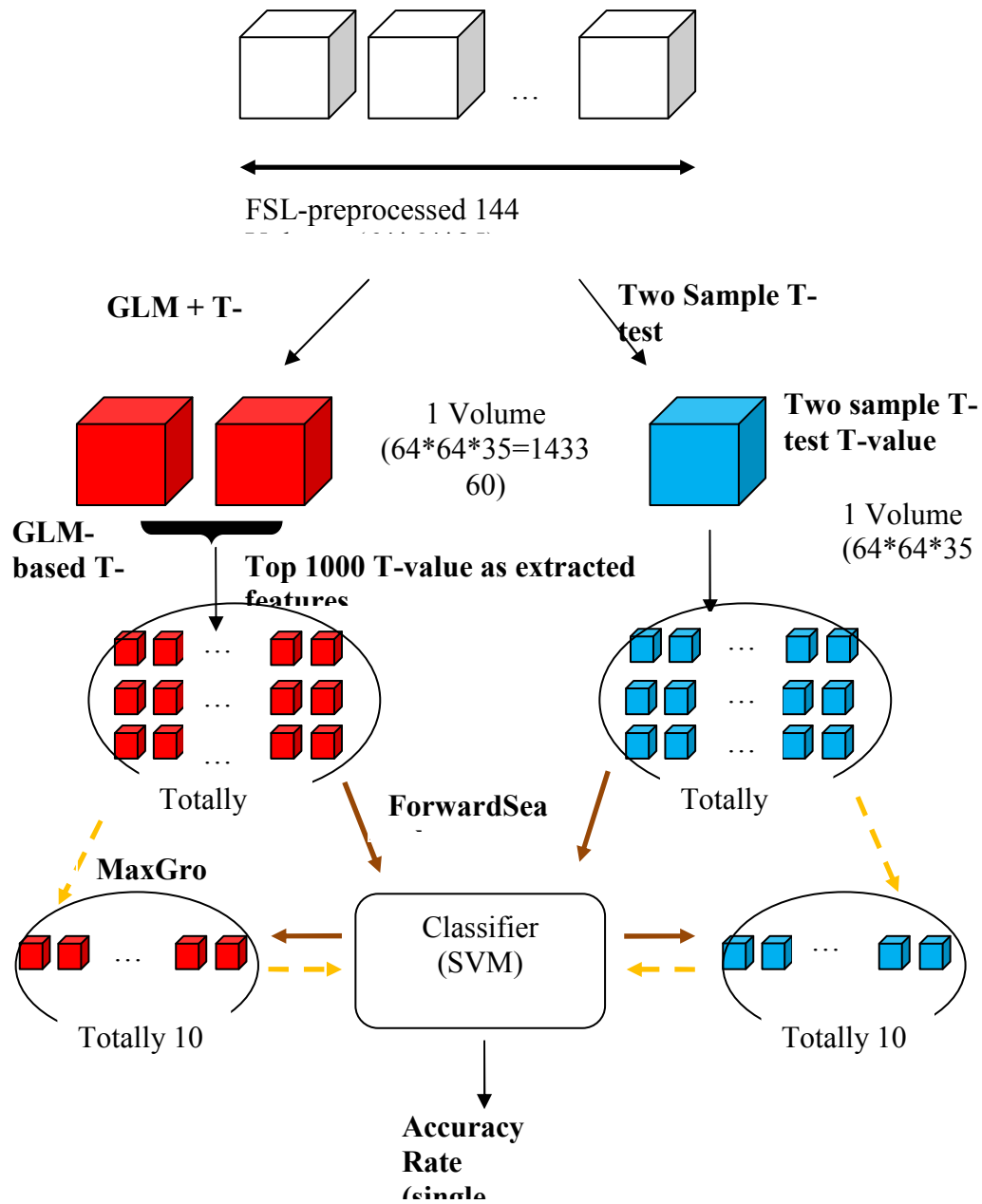


Figure 9 Main steps of the whole algorithm

The figure 9 above describes the main steps of the whole algorithm. Subject3 will be used to describe the blocks in this algorithm.

SUBJECT-3 4D fMRI image contains: 144 volumes, each volume's size is 64*64*35.

Ten features will be selected from extracted features (by PCA) as input of classifier.

3.4 Reorganization of Data in PCA Input Form

The raw data is 4D, which is $64*64*35*144$. We take the 144 volumes as 144 “observations”, and each observation contains 143360 voxels (one volume’s voxels). These 143360 voxels can be viewed as “original, high-dimensional coordinate system” whose dimension is 144360. Note that now the data has been transformed from 4D to 2D.

3.5 Feature Extraction – PCA

PCA can compress the dimension of images by projecting the data (144 observations) from original high-dimensional coordinate (144360-D) to a new, low-dimensional coordinate system. After PCA, two matrices can be obtained in MATLAB, one is $143360*143$, named “COEFF”, and the other one is $144*143$, named “SCORE”. It means that *PCA-processed SUBJECT-3* image has been projected to a new, low-dimensional coordinate system (143-D); each of this new coordinate’s 143 indices is formed by all 143360 indices of original, high-dimensional coordinate (that is, “COEFF”). *PCA-processed Subject1* image still have 144 observations, but the data has been represented in this new 144-D space (that is “SCORE”). To perform the pattern recognition, only the matrix “SCORE” is needed.

3.6 Feature Selection-Max/ForwardSearch

Max-group method simply takes the top 10 values in the SCORE (that is, columns 1 to columns 10) as features. In PCA theory, they represent values got from projecting data into 10 most significant indices of “new, low-dimensional coordinate system”, which contain the most part of variance in the data.

Forward-Search method tests, chooses and adds the best performance’s feature to the feature group. It cooperates with classifier. First, it tests all features one by one with classifier to find the best one and form the one-feature group. Then it tests all features

except the chosen one again and finds another best one to form the two-feature group. Then it tries to find the third one to form a three-feature group and so on. The algorithm keeps running until 10 features found. It is a little different from the standard Forward-Search algorithm because this modification can help to avoid the local extrema in the search procedure.

3.7 Training and Testing the Classifier

K-fold validation is used to train and test the classifier. K-fold validation is used to get reliable performance evaluation of one classifier. For supervised-classification learning algorithm, our dataset is divided into train-set and test-set. However, the performances obtained from training and testing our classifier one time or several times are not reliable. So all the data is needed to train and test the classifier and the performances are needed to be averaged to eliminate the outlier (that is, too high or too low accuracy rate got by chance).

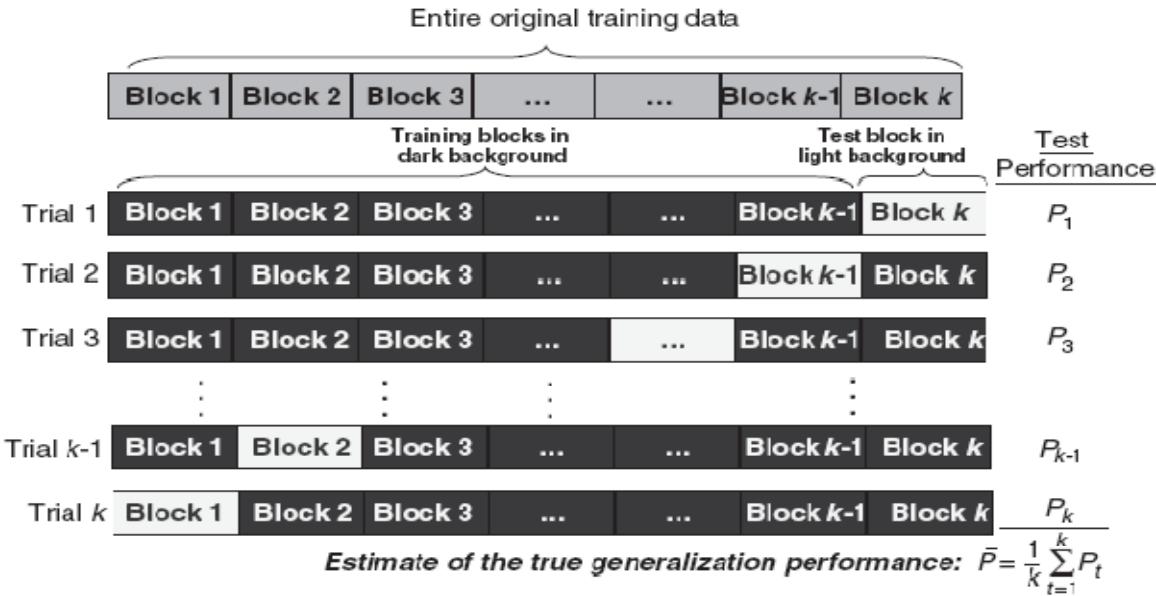


Figure 10 k-fold validation

The figure 3 above describes the k-fold validation. The whole data has been divided into k folds. One fold is used to test and the rest ones are used to train, namely “one trial”. Totally

k trails are needed to be performed to ensure that in each trail a different fold is used as test-fold. After each trail, one performance (that is, accuracy rate of that trail) is obtained and after all k trails finished, the average of those k trail-performances is calculated and this average is taken as “estimate of the true generalization performance.”

In our case, *SUBJECT-3*'s data has 144 volumes. They include 54 counting volumes, 54 reasoning volumes and 36 rest volumes. Among them, 54 counting volumes and 54 reasoning volumes are needed. It means that 108 samples will be used, in which half of them are counting samples and others are reasoning samples.

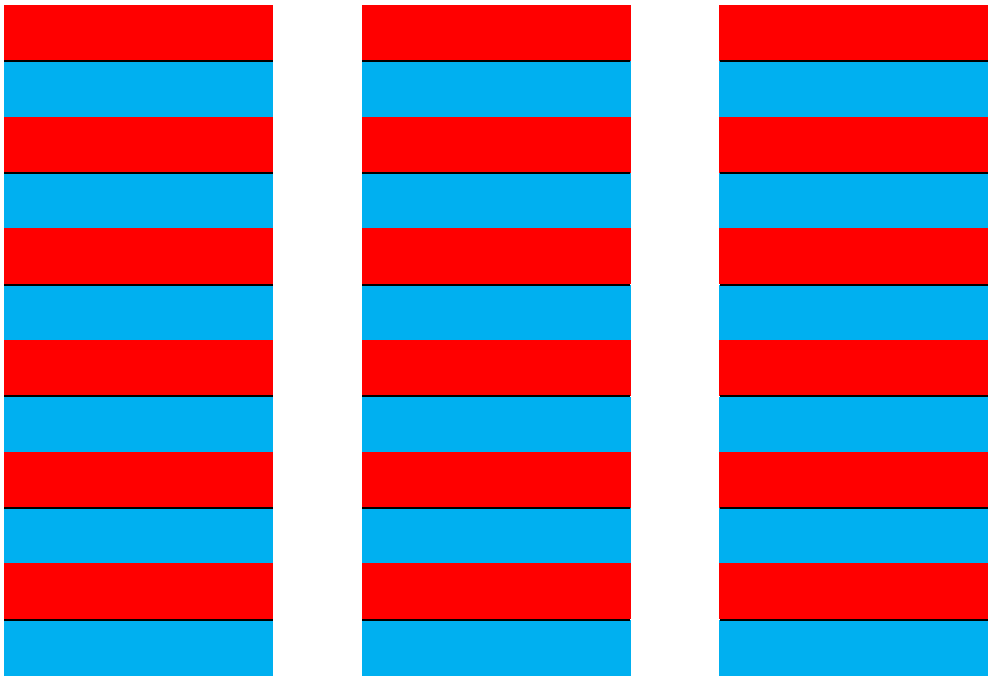


Figure 11 Constitution of 144 volumes (Red = Counting, Blue = Reasoning, White = Rest)

In our case, k is set to 12. It means that the whole useful data (that is, 108 counting/reasoning volumes) is divided into 12 folds, and each fold contains 9 volumes. Training will use 11 folds (totally 99 volumes), and the rest one fold (totally 9 volumes) will be used to test the trained classifier. Each fold of 12 folds will be used as the test fold once. So 12 training and testing with different 12 test folds will be performed. The final accuracy rate will be the average of these 12 tests.

$$\text{Accuracy Rate} = \frac{1}{12} \sum_{i=1}^{12} (\text{Accuracy Rate})_{i\text{-th test}}$$

3.8 Feature Extraction – GLM-based t-test/Two-sample t-test

GLM-based t-test is performed by software FSL (FMRIB Software Library). As *Figure 4* shows, standard routine is performed, which contains four different contrasts (Counting only (C_Only), Reasoning only (R_Only), Counting's response is stronger than reasoning (C>R) and Reasoning's response is weaker than reasoning (C<R)). However, only contrast 3 (C3: C>R) and contrast 4 (C4: C<R) will be used in this report.

FSL uses all default settings in the experiment. Key parameters are listed below:

Data tab:

TR (s) = 3.0

High pass filter cutoff (s) = 100

Pre-stats tab:

Spatial smoothing FWHM (mm) = 5

FSL finishes all the work and provides the t-test result with two .nii files (*/stats/tstat3.nii* and */stats/tstat4.nii*). It is noted that the result from FSL t-test includes both positive and negative values. It is because that FSL use double-side t-test. All t-test values need to be transformed to be positive via taking the absolute value of negative ones.

Two-sample t-test is mainly based on the equation (**Eq.1**). The computation is performed one voxel by another. The result is similar as the GLM-based t-test produced by FSL, which contains both positive and negative values. All negative values need to be changed to their absolute values.

After GLM-based t-test, two volumes (one for contrast 3 and the other for contrast 4) which contain t-values are provided. Top 500 voxels are chosen from each volume (totally, 500*2

= 1000) as the extracted features. For Two-sample t-test, only one volume is provided, so top 1000 voxels are chosen as extracted features directly. **Max-group method** simply takes the top 10 values in the extracted features. **Forward-Search method** tests, chooses and adds the best performance's feature to the feature group. It cooperates with classifier. More details can be found in the report "Pattern recognition of fMRI image using PCA method"

4. Results

There are 8 different subjects used in our report. Two different feature extraction methods are used: GLM-based t-test and Two-sample t-test. After 1000 features extracted, both Max-group and Forward-Search feature selection methods are used to get 10 features as the input of our classifier.

The classifier is Support Vector Machine (SVM). It is from *MATLAB bioinfo toolbox*. All parameters of SVM are set to their default values.

4.1 Extracted features from different feature extraction methods

The GLM-based t-test feature extraction method and Two-sample t-test feature extraction method produce different extracted features from original data. The following example is the result of subject *BRANDON*. Both results from GLM-based t-test feature extraction method and Two-sample t-test feature extraction method have been marked on that subject's fMRI image with different colors. Red points represent results from GLM-based t-test feature extraction method and green points are the results from Two-sample t-test feature extraction method.

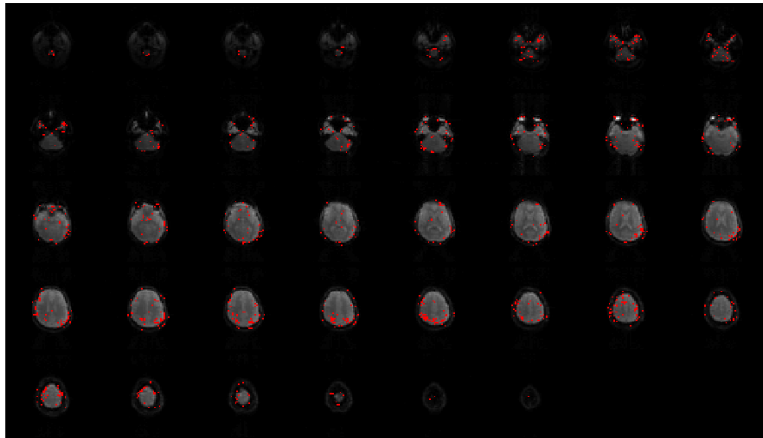


Figure 12 Results from GLM-based t-test feature extraction method

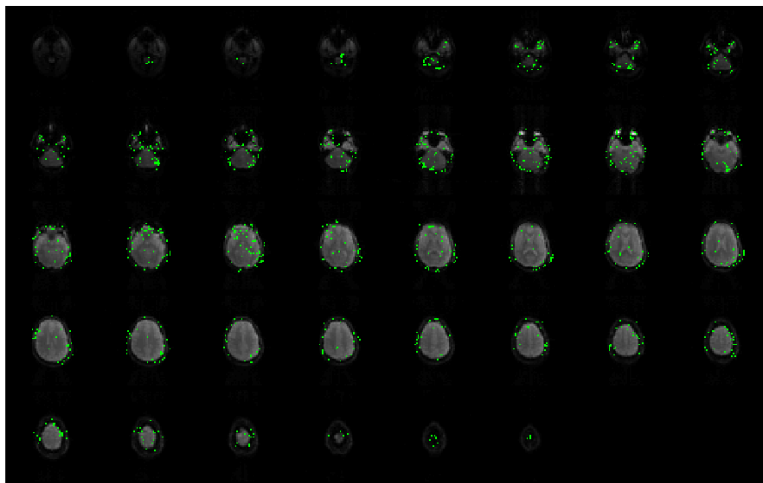


Figure 13 Results from Two-sample t-test feature extraction method

4.2 Selected features from different feature selection methods

After feature extraction, Maxgoup and ForwardSearch Feature Selection methods form the different selected-feature group (each contains 10). There are 4 different combinations of feature extraction and feature extraction methods to get 4 different selected-feature groups. They are: GLM-Max group, GLM-Forward group, Two-sample-Max group and Two-sample-Forward group. They are demonstrated with different color using the results of subject *BRANDON*. Red points represent GLM-Max group, green GLM-Forward group, blue Two-sample-Max group and yellow Two-sample-Forward group.

It is noted that all these images below has been processed with “eight-neighbor enhancement” to improve its effect (or it is really difficult for many people to find independent pixel in the image), which means that one color point (in fact, it is a square) on the image contains feature itself and its 8 neighbors.

Figure 7 describes the principle of “eight-neighbor-enhancement”

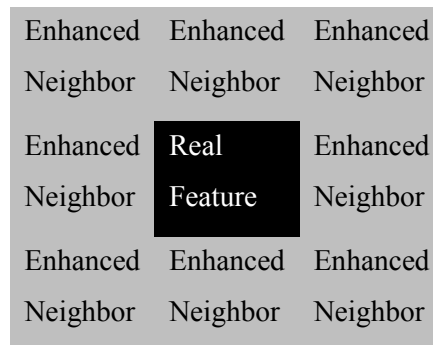


Figure 14 Eight-neighbor Enhancement used in selected-feature demonstration

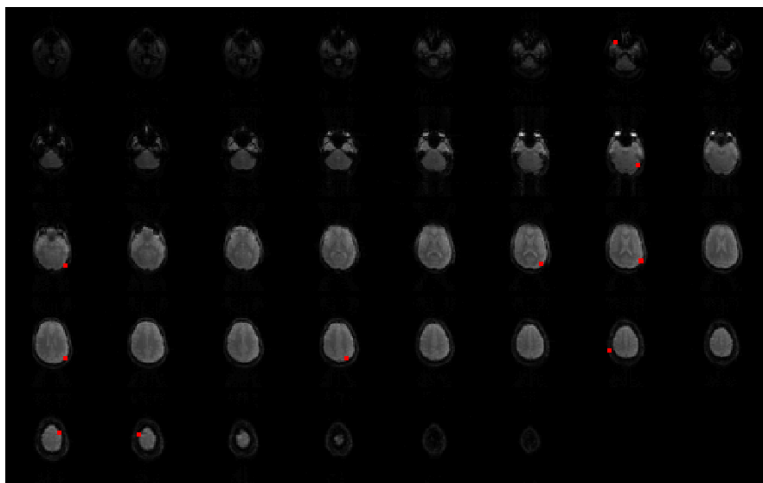


Figure 15. Ten selected-features in GLM-Max group

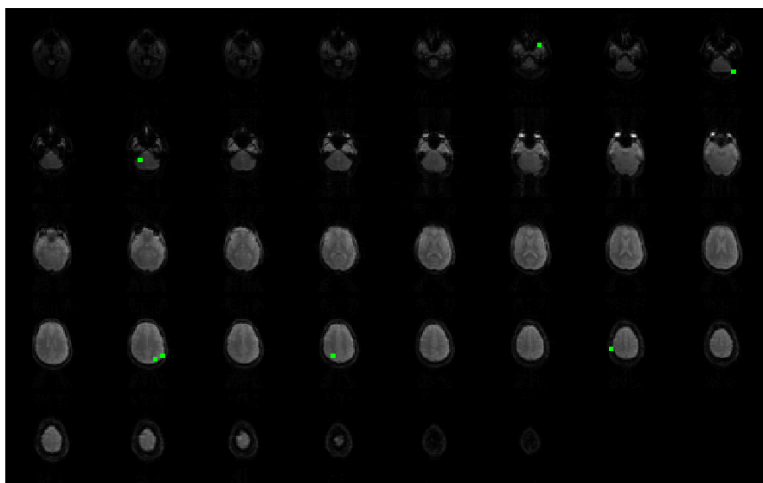


Figure 16. Ten selected-features in GLM-Forward group

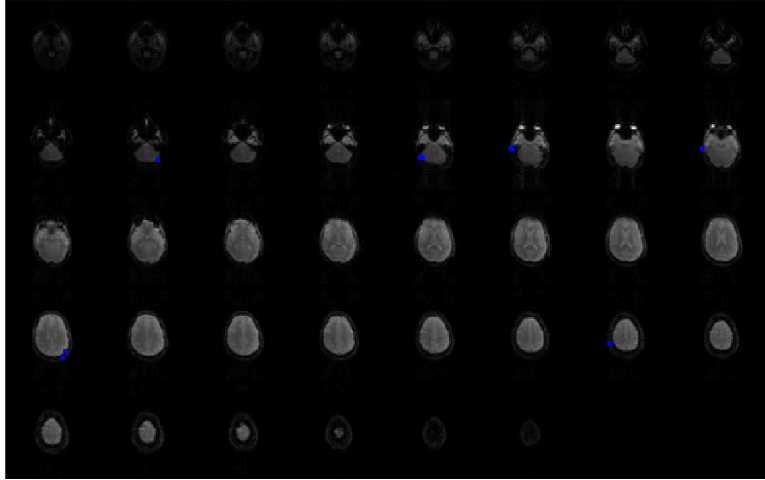


Figure 17 Ten selected-features in Two-sample-Max group

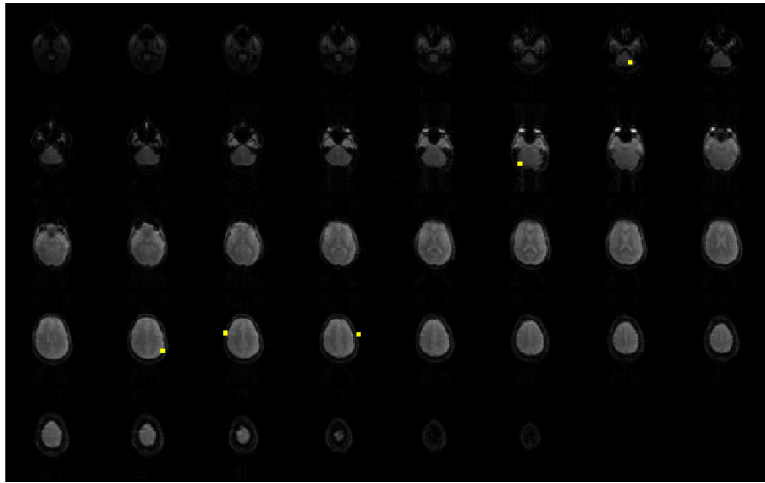


Figure 18 Ten selected-features in Two-sample-Forward group

4.3 Accuracy rate from different feature extraction and selection methods

The accuracy rate of 8 subjects' test using SVM with different feature extraction and selection methods is below:

	GLM/Max	GLM/ForwardSearch	Two-sample/Max	Two-sample/ForwardSearch
SUBJECT-1	66.7	76.9	79.6	89.8
SUBJECT-2	72.2	83.3	82.4	85.2
SUBJECT-3	74.1	87.0	83.3	90.7
SUBJECT-4	71.3	78.7	83.3	89.8
SUBJECT-5	67.6	83.3	81.5	82.4
SUBJECT-6	64.8	82.4	80.6	84.3
SUBJECT-7	69.4	74.1	70.4	72.2
SUBJECT-8	57.4	82.4	77.8	90.7

Table 1 Accuracy rate (percentage) of 8 subjects using SVM and different feature extraction and selection methods

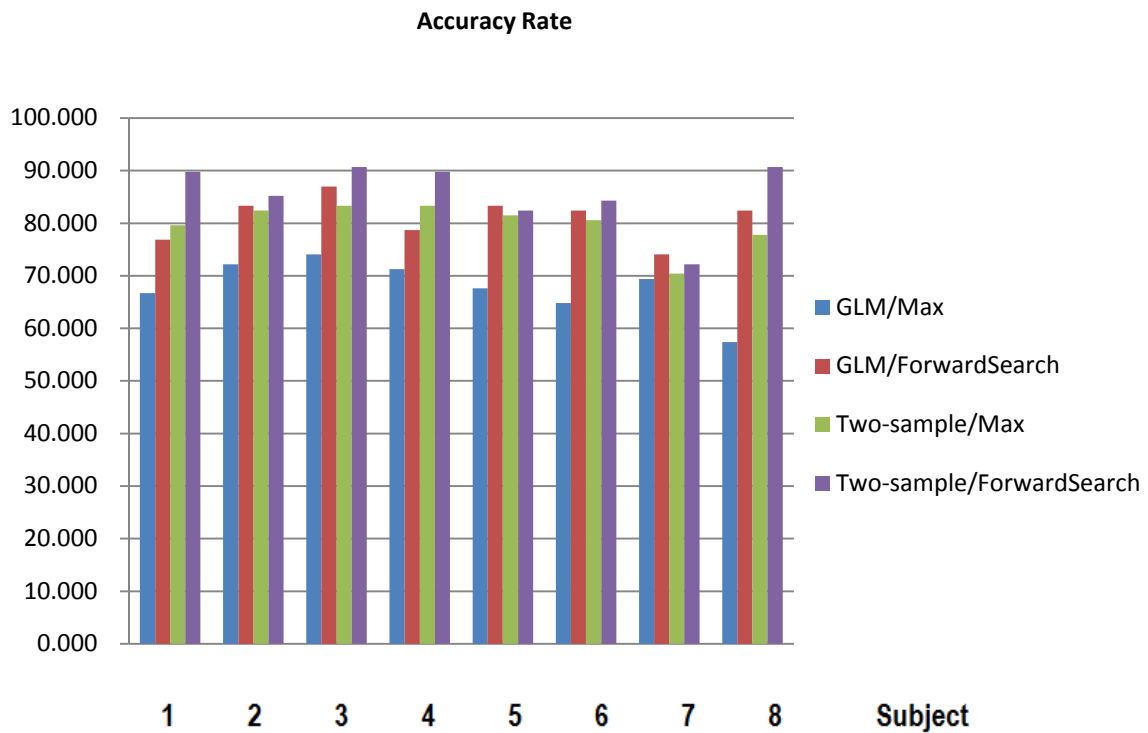


Figure 19. Comparison of accuracy rate between different feature extraction and selection method

4.4 Results on classification

There are 8 different subjects used in our report. Two kinds of FMRI image source of each subject are used: filtered (which includes the motion-correction and high-pass filtering) and unfiltered (that is, the raw data). After PCA-based feature extraction, both Max-group and Forward-Search feature selection methods are used to get 10 features as the input of our classifier.

Three classifiers are used independently. They are Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (KNN). Both SVM and KNN are from *MATLAB bioinfo toolbox*. LDA is from *MATLAB statistics toolbox*. All parameters of three classifiers are set to their default values.

The result of 8 subjects' test using SVM is below:

	Unfiltered/Max	Unfiltered/ForwardSearch	Filtered/Max	Filtered/ForwardSearch
SUBJECT-1	25.0	68.5	45.4	75.9
SUBJECT-2	35.2	71.3	48.1	68.5
SUBJECT-3	43.5	64.8	54.6	75.9
SUBJECT-5	27.8	68.5	48.1	79.6
SUBJECT-6	29.6	72.2	50.0	74.1
SUBJECT-7	58.3	67.7	55.6	76.9
SUBJECT-8	51.9	71.3	56.5	70.4

Table 1 Accuracy rate (percentage) of 8 subjects using SVM

5. Conclusion

Among three classifiers, with the same feature-selection method and source, SVM and LDA can provide better performance than KNN. It is probably caused by “active” characteristic of SVM and LDA. SVM and LDA can project the data into a “new data space” to make classification easier. KNN is nearly “passive”, which means it relies on data’s own distribution character more than trying to make classification easier via transformation. Generally speaking, SVM should provide better performance than LDA. However, in our case, SVM and LDA can provide similar performance. It can be explained that as PCA projects the data to a new space “which can reflect most part of the variance of the data”, the PCA-processed data is “scattered”. In our case, the degree of scatter is large enough to make LDA achieve the performance which is comparable with SVM’s. Under the condition of using the same feature-selection method, filtered image can provide better classification performance than unfiltered image. With the same image source, Forward-Search feature selection can provide much better classification performance than Max-group feature selection, especially when unfiltered images are used. However, Forward-Search selection is slower than Max-group selection. There is a tradeoff between speed and accuracy rate when you choose feature selection method. In practice, different application may emphasize speed or accuracy. Different emphasis will produce different choice on feature selection methods.

Generally speaking, with Two-sample t-test feature extraction and Forward-Search feature selection method, SVM classifier can provide about 90% accuracy rate. It is much higher than the result from transformation-based feature extraction method, for example, PCA-based feature extraction method can reach only 70% in the last report “Pattern recognition of fMRI image using PCA method”. Under the condition of using the same feature-selection method, Two-sample t-test feature extraction can provide better classification performance than GLM-based t-test feature extraction. With the same feature-extraction method, Forward-Search feature selection can provide much better classification performance than Max-group feature selection. However, Forward-Search selection is slower than Max-group selection. There is still a tradeoff between time and accuracy in this case.

It needs to point out that the tradeoff in this report is quite different from the time-and-accuracy tradeoff problem in the last report “Pattern recognition of fMRI image using PCA method”. Since the extracted features increase from 143 (PCA-extraction) to 1000, Forward-Search in this report is much less attractive than it was in the last report because it increases accuracy rate little but costs much more time (Two-sample/Max (80%, 15mins, 8 subjects) vs. Two-sample/ForwardSearch (90%, 5hours,8 subjects)). So among all the combinations of feature-extraction methods (PCA/GLM-based t-test/Two-sample t-test) and feature-selection methods (MaxGroup/ForwardSearch) discussed in this report and the last report (“Pattern recognition of fMRI image using PCA method”), “Two-sample t-test + ForwardSearch” can provide the highest accuracy rate. However, with consideration of time spent in the whole procedure, “Two-sample t-test + MaxGroup” seems to be the best choice.

Acknowledgments

The authors would like to thank Dr. Michael O’Boyle for providing the data for the experiments reported in this work.

The first author also acknowledges the support received from Dr. Brian Nutter and Dr. Sunanda Mitra, CV&IA Lab, and the Electronics Department, INAOE, during the research stay.

References

1. M.W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, S.M. Smith. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45:S173-186, 2009.
2. S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J.M. Brady, and P.M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):208-219, 2004.
3. Doug Greve, "Statistics Review for fMRI Data Analysis," *HST.583: Functional Magnetic Resonance Imaging*, pp. 10-11.
4. Serdar Kemal Balci, "Classification of whole brain fMRI activation pattern," *Degree thesis for Master of Science (MIT)*, 2008, p. 32.
5. Stephen Smith D.G. Leibovici, "Comparing groups of subjects in fMRI studies: a review of the GLM approach," *Center for Functional Magnetic Resonance Imaging of the Brain, University of Oxford*, 2001, p. 3.
6. Francisco Pereira, Tom Mitchell, Matthew Botvinick, "Machine learning classifiers and fMRI: a tutorial overview"
7. Martin J McKeown, Lars Kai Hansen, Terrence J. Sejnowski "Independent component analysis of functional MRI: what is signal and what is noise?", *Neurobiology* 2003, Vol. 13, pp. 620–629.
8. C. Davatzikos, K. Ruparel, Y. Fan, D.G. Shen, M. Acharyya, J.W. Loughhead, R.C. Gur, D.D. Langleben, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection", *NeuroImage*, Vol. 28, 2005 pp. 663 – 668.
9. David D. Coxa,b,* and Robert L. Savoy, "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex", *NeuroImage*, 2003.
10. Kriegeskorte, N., Goebel, R. & Bandettini, P. "Information-based functional brain mapping", *Proc. Natl Acad. Sci. USA* 103, 3863–3868, (2006).
11. Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H. & Stetter, M. "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data", *Neuroimage* 28, 980–995 (2005).

12. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. “Intersubject synchronization of cortical activity during natural vision”, *Science* 303, 1634–1640 (2004).
13. Saleem Zaroubi, Gadi Goelman, “Complex denoising of MR data via wavelet analysis:
14. Application for functional MRI”, *Magnetic Resonance Imaging* 18 (2000) pp. 59–68.
15. Alice J. O’Toole, Fang Jiang, Herve Abdi, Nils Penard, Joseph P. Dunlop, Marc A. Parent, “Theoretical, Statistical, and Practical Perspectives on Pattern-based Classification Approaches to the Analysis of Functional Neuroimaging Data”, *Journal of Cognitive Neuroscience* 19:11, pp. 1735–1752, 2007.
16. M. Jenkinson and S.M. Smith. “A Global Optimisation Method for Robust Affine Registration of Brain Images. *Medical Image Analysis*”, 5:2(143-156) 2001.
17. M. jenkinson, P. Bannister, M. Brady and S. Smith. “Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images”, *NeuroImage* 17:2(825-841) 2002.
18. Serdar Kemal Balci, “Classification of Whole Brain fMRI Activation Patterns”, 2008
19. Minh-Tuan T.Hoang “Cognitive States Detection in fMRI using incremental PCA”, 2007
20. ICA in FMRI: A Basic Introduction (report), Mark Jenkinson, 2003
21. Mitchell, T.M., Classifying instantaneous cognitive states from FMRI data. *AMIA Annu. Symp. Proc.* 465–469 (2003).

Appendix A Principal Component Analysis

Principal Component Analysis (PCA) is used to decrease the dimension of data and extract the most important features of that data for pattern recognition.

PCA has the assumption that the variance of data contains the information we need. So PCA tries to get the as much variance as possible in data. PCA view the original high-dimensional coordinate system as “redundant”, so PCA tries to find a transformation R which can make new representation (in new, low-dimensional coordinate system) C contains the most part of variance in data X and at the same time, needs as less dimension (that is, number of indices of coordinate system) as possible. It means that PCA tries to find a transformation R having $R^T(XX^T)R = \Lambda^5$. Generally, R can be found by eigenvector decomposition or SVD (Singular value decomposition)⁶. After finding R , X will be represented by C via $C = R^TX$.

Typically, just using first several elements (generally, columns) of R is enough to represent or reconstruct the data, so these elements are called “principal component”.

Appendix B Support Vector Machine

Support Vector Machine (SVM) is used for two-class classification. SVM tries to find a vector w which makes samples x from two different classes (class +1 and class -1) can be divided by two hyperplanes (plane +1 and plane -1). These two hyperplanes have the maximum distance between each other (that is, maximum margin).

SVM uses “quadratic programming”⁷ to find this w , that is:

Minimize

$$\frac{1}{2} w \cdot w + C \sum_{k=1}^R \varepsilon_k \quad (\text{R is number of samples, } \varepsilon_k \text{ is the distances of error points to their correct plane)}$$

With constrain

$$w \cdot x_k + b \geq 1 - \varepsilon_k \text{ if } y_k = 1$$

$$w \cdot x_k + b \geq -1 + \varepsilon_k \text{ if } y_k = -1$$

Sometimes, SVM needs to use some kernel to rearrange x to find w more easily. The simplest kernel is linear kernel. There are several other kernels, such as quadratic kernel, Gaussian Radial Basis Function kernel and polynomial kernel.

Appendix C Linear Discriminant Analysis

Linear Discriminant Analysis tries to find a transformation which can make intra-class distance minimized and inter-class distance maximized.

LDA define the intra-class distance as within scatter matrix S_W and inter-class distance as between scatter matrix S_B ⁸. Similarly as PCA, LDA tries to find a transformation W to project data into a new space via $y = W^T x$.

$$S_W = \sum_{i=1}^C \sum_{x \in X_i} (x - m_i) \cdot (x - m_i)^T \quad ^9$$

$$S_B = \sum_{i=1}^C n_i (m - m_i) \cdot (m - m_i)^T \quad ^9$$

C is the number of classes, x_i is samples in i th class. m_i is the mean of all samples in i th class. m is mean of all samples in all classes. n_i is number of samples in i th class.

Elements of w (generally, columns) can be got from the eigenvectors corresponding to largest eigenvalues in the following equations⁹:

$$S_B w_i = \lambda_i S_W w_i$$

Appendix D K-Nearest Neighbor

K nearest neighbor (KNN) is one of the simplest classification methods. It needs little training about classifier. To do classification, KNN just needs data whose label (that is, the class of the data) is known and the definition of distance. There are two kinds of distances which are used most commonly: Euclidian Distance and Mahalanobis Distance. When KNN starts to work, it accepts a new sample which label (that is, class) is unknown, then KNN will calculate the predefined distance between this sample and elements of label-known data. KNN will find the K smallest distance among all these distances and call K elements from label-known data corresponding to those K distances as K “neighbors” of that label-unknown sample. Generally, these K neighbors come from more than one class, so KNN will assign the class, which has most elements in these K neighbors, to that unknown sample.