fMRI Brain Image Retrieval Based on ICA Components

Bing Bai
bbai@cs.rutgers.edu
Department of Computer Science
Rutgers University
Piscataway, NJ 08854

Paul Kantor
kantor@scils.rutgers.edu
Department of Library and Information Science
Rutgers University
New Brunswick, NJ 08901

Ali Shokoufandeh
ashokouf@cs.drexel.edu
Department of Computer Science
Drexel University
Philadelphia, PA 19104

Deborah Silver
silver@ece.rutgers.edu
Department of Electrical and Computer Engineering
Rutgers University
Piscataway, NJ 08854

Abstract

This manuscript proposes a retrieval system for fMRI brain images. Our goal is to find a similarity-metric to enable us to support queries for "similar tasks" for retrieval on a large collection of brain experiments. The system uses a novel similarity measure between the result of probabilistic independent component analysis (PICA) of brain images. Specifically, the times series of an fMRI dataset will be represented using a number of ICA components as high level task-related features. The similarity between two datasets is the value of the maximum weight bipartite matching defined on the component-wise similarities. The component-wise similarities are calculated based on the size of the overlap between the "highly activated" regions in the corresponding activation maps. We evaluated the performance of the proposed method on a moderate size fMRI image database with considerable variety. The ICA-based component selection in combination with bipartite matching similarity measure outperforms several other component selection methods and similarity measurements. The results also suggest that there is a direct correlation between the involvement of ICA components in cognitive processes and their time course spectrum. Along with other heuristics, this property can be for fMRI image retrieval and classification.

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is a technique used to "monitor" brain activities [10]. The most widely used fMRI method is BOLD (blood oxygen level dependent) fMRI, in which the intensity of an image element (it is called a *voxel* in 3D images, corresponding to a "pixel" for 2D images) is related to the level of blood oxygen in the corresponding region. When a cognitive process involves a specific brain region, at first, some oxygen is consumed, but more oxygen is brought in by blood flow soon after, and this change will appear as increased intensity in the corresponding region of the image. In a typical fMRI experiment, the experimental subject is assigned a certain type of cognitive task (which is referred to here as the *stimulus*) at some scheduled moments, while, between tasks, the subject focuses on something different from the task, and relatively undemanding (like watching a cross hair). A 3D brain scan is made every few seconds, during each experimental session. By comparing the intensities of voxels during the condition and the control periods, we can estimate which brain regions are "activated" by the cognitive process. In this paper, we refer to a sequence of 3D images from a experimental session as a *dataset*. Note that it may have different meanings in other fMRI papers.

As a method for watching "how the brain works", fMRI has been used as a powerful research tool in many (healthy) neuroscience studies in the past decade, and it is also gaining clinical attention more recently. For example, studies [21] of Alzheimer's disease show that differences from a typical brain can be detected using

scans before the symptoms are otherwise apparent, for some Alzheimer's patients.

Many fMRI experiments have been conducted, and we expect many more of them in future. As the size of the (world) data corpus increases, efficient data *sharing* schemes for fMRI data become more and more desirable. Data sharing, of course, provides a larger database for testing and validating analytical algorithms under development. More importantly, however, different researchers may find different value in the same data, discovering similarities in the brain's activity, when the cognitive tasks do not seem to be related, based on psychological reasoning alone.

In this paper, we report some investigations of *content-based* retrieval of fMRI images by using *task* labels. In other words, when a "query" fMRI image is presented, we ask whether we can return images that represent the same or similar cognitive processes. The potential applications include, but are not limited to, the following: (1) Helping researchers find similar studies and related research work. (2) Helping researchers discover hidden similarities among superficially different studies.

It is straightforward to map the brain image retrieval problem to the framework of content-based image or video retrieval. However, all fMRI images have similar brain structure, similar shape and similar color/intensity range. Thus, the common features used in the image retrieval community – such as colors, textures, and shapes – can not be applied to fMRI data directly. In particular, all fMRI images have similar brain structure, similar shape and similar color/intensity range. Thus, in indexing fMRI images, we are looking for very subtle intensity changes over space or time, which is not the goal of most image/video retrieval algorithms.

Some related work addresses "classification" of brain images. Some researchers try to detect "activation" volumes in the same brain image sequence [18, 15], while others try to distinguish experiments with different cognitive tasks [8]. In these studies, machine learning (ML) methods (k-nearest neighbors [18], Bayesian [18], Support Vector Machines (SVM) [18, 15] and Fisher's Discriminant Analysis [8]) are applied using features such as the time series of voxels [18, 15], or to intermediate result from other processing, such as t-maps [8] generated by General Linear Model (GLM) (see next section for details of GLM). All ML methods characterize the distribution of some set of features for labeled training datasets, and use these characteristics to classify other datasets.

In recent few years, a number of papers has been published on content-based fMRI retrieval [2, 3, 24].

These methods are mostly based on GLM t-maps, which require accurate information of stimulus time series and of hemodynamic response models.

In this paper, we propose a similarity metric for brain images based on independent component analysis (ICA) [14, 17, 4] (see next section for details of ICA). The advantages of ICA are that: (1) It can be applied on datasets whose stimulus is unknown or is not well defined (e.g., some fMRI experiments let subjects watch a episode of movie).(2) Even if the stimulus is known, the individual hemodynamic response may be different from the assumed models. Thus is potentially valuable to discover a model instead of apply a model.

Similar to PCA, ICA decompose the observations into latent components. However, unlike PCA, which decorrelate the components, ICA pursues statistically independent components. In the context of fMRI images, ICA can isolate factors which have physiological and psychological meanings, instead of pure dimension reduction. Thus, ICA could be a very good feature extraction for brain image classification by the cognitive tasks, if the task related components can be identified. In [13], the components are considered more task related if their time courses have stronger correlation to known stimulus time series. However, for datasets with more than one cognitive tasks, it may be difficult to tag the independent components.

Instead of matching the components' time courses with stimulus, we propose a more flexible, yet more stable heuristic. We select a number of components whose energy concentrates on the low end of the frequency spectrum. These similarity between datasets is based on the "best" matching, which is represented by the maximum weight bipartite matching (MWB) [5], of these feature components. We tested the proposed method, in a retrieval framework, with a database of 360 datasets from 8 different tasks. Our result shows that the heuristic of low expected frequency with MWB matching outperforms other methods.

Our contributions can be summed as the follows: (1) We presented a retrieval framework based on ICA components. To the best of our knowledge, ICA has not been applied to large data repository retrieval yet. (2) We introduced "low frequency" heuristic for selecting task-related components. (3) We proposed to use MWB matching for image similarity. (2) and (3) together gives best performance in our tested methods.

2 Background

2.1 The General Linear Model (GLM)

The most widely used mathematical method for fMRI time series analysis is the general linear model (GLM) [11, 9]. In the GLM, the time series of intensity value at each voxel is modeled as a linear combination of explanatory variables and Gaussian noise. An explanatory variable arises from the hypothesized response to a certain type of stimulus. Because it takes a few seconds for the new supply of oxygen to reach an active region of the brain, each explanatory variable is generated by convolving the stimulus time series with a specific impulse response function, which is referred to as hemodynamic response function (HRF). The mean and variance of the weight (in the regression) of each explanatory variable is calculated by linear regression. By comparing each weight with its estimated variance one can estimate the (Fisher) t-statistic, which will be referred as the t-value of the weight. It is a measure of the degree to which the level of activation is influenced by the corresponding stimulus.

2.2 Independent Component Analysis (ICA)

While the GLM approach specifically considers the time variation of the stimulus, in identifying voxels "of interest", another important method is independent component analysis (ICA). There are a lot of publications that we can not cover in this paper, please see [14] for an overview. Noiseless ICA models fMRI signals in terms of underlying or latent components: $X = AS$, in which X is the observation matrix, in which columns are time series, and rows are voxels in a brain. S is the assumed signals. A is the mixing matrix. However, ICA does not assume the knowledge of either the stimulus or the hemodynamic response function. Instead, ICA postulates that the rows in S are independent, and estimates the linear transformation A that maximizes the independence between components of S. Each row of S can be viewed as a activation map of a time course that was presented in the corresponding column of A.

ICA was introduced into fMRI community by [17]. Considerable work has been done after that, please see [6] for review. Probabilistic ICA (PICA) [4] proposed a "noisy" ICA model, which is similar to general linear model:

$$X = AS + \eta$$

in which η is Gaussian noise. Instead of making A a full rank square matrix, the dimension is reduced first with probabilistic PCA (PPCA) [22]. Thus, X can not be fully explained as a linear combination of the rows in S like in noiseless ICA model. The residual is considered as noise to do statistical inference. An ICA component includes a time course (which is a column in A) and a brain map in which each voxel is a z-value that indicates the activation of the voxel. This statistical map is referred to as *z-map* in this note. PICA is implemented in fMRI analytical software FSL [20] as "melodic" module, and it is the ICA model we use in this paper.

3 Methods

3.1 Preprocessing and Registration

The raw fMRI images from different scanners, or for different subjects are not comparable. Raw signals contain many kinds of noise. These noises should be controlled by following preprocessing steps before ICA is applied: (1) Apply motion correction to align all 3d images in a time series to the same position, to correct the effects caused by small movements of the subject during the experimental run. (2) Remove the skull since it is not a part of the brain. (3) Apply a temporal high pass filter to remove low frequency trends over time due, for example, to increasing temperature of the device. (4) (Optional) Apply spatial smoothing filters to control spatial noise.

Another important issue is registration. Different brains have the same structures, but their shapes and sizes may vary. Registration is the operation that transforms brain images onto a standard brain template to allow inter-subject comparisons. In this work, we choose to do registration after ICA. The "standard space" has smaller voxels, and thus building activation maps would take longer in the standard space. However, the apparently larger number of voxels does not really represent more information, since they are calculated from the smaller number of voxels in the raw image. Thus there is no loss in doing voxel selection before the transformation. We conducted preprocessing and registration using the fMRI software package FSL.

3.2 Image Similarities

We know that PICA decomposes a temporal-spatial matrix into "independent components" (independent spatial maps and corresponding time courses). In the PICA results for our datasets, there are usually 20 to 150 components for each dataset, after the dimension reduction of PPCA. Some of these components, hopefully, are task-related hemodynamic responses, while
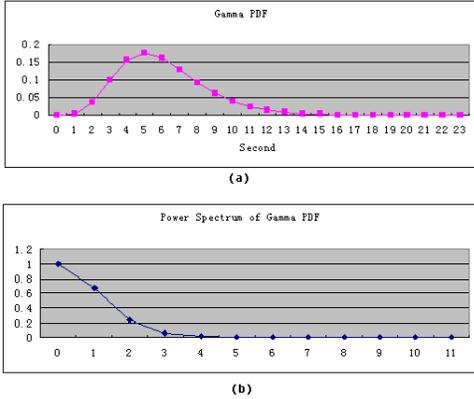
Figure 1. (a) Gamma function. (b) The power spectrum of the Gamma function

Figure 2. The process to calculate similarity between datasets.

others are all kinds of physiological or psychological noise [6]. Here we propose a simple heuristic based on the spectrum of the time course. In classical fMRI theory, a task related time course (also called *Explanatory variable*) is modeled with the convolution of HRF and the stimulus time series. While there are debates about the shape of the HRF, it is often modeled as a Gamma PDF (Figure 1), which reaches the peak in around 6 seconds, and decays after that. In Figure 1 we see that the mass of energy for a Gamma PDF is on the low end of the spectrum, implying that the Gamma PDF is a low-pass filter. By the convolution theorem:

$$F(f * g) = F(f) \cdot F(g),$$

we know that the energy of explanatory variable must also tend to stay in the relatively low end of the spectrum. We use the following metric, which we call *expected frequency* to indicate the energy tendency:

$$E(f) = \frac{\sum_{i=1}^{N} i \cdot A_i}{\sum_{i=1}^{N} A_i}$$

A_i is the amplitude of ith element in the discrete power spectrum. To allow some flexibility, we choose a number of components with the lowest expected frequency from each dataset. These are referred to as "feature components" in the rest of this paper.

The similarity between two datasets is derived from the similarity of their feature components. We do not have prior knowledge on how these components correspond to each other, so we have to test each component in one dataset against every component in the other dataset (the similarity measure between two components is addressed in next subsection). Now we have a complete weighted bipartite graph, of which each edge
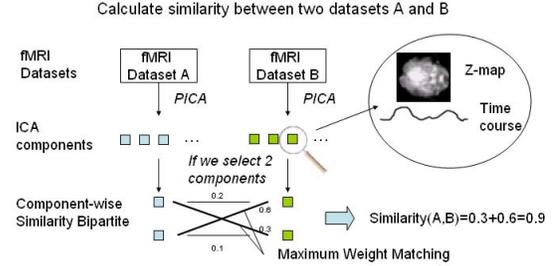
has a weight representing the similarity score between these two components. We then find the Maximum Weight Bipartite (MGB) matching [5] in this bipartite graph, and use the sum of the weights as the similarity between the two datasets. MGB has already been used in image matching applications (e.g. [7]). Figure 2 is a diagram for the above process:

3.3 Component Similarities

As noted, each PICA component is a statistical z-map and it is associated with a time course. We can not get any information by comparing time courses, because the task related time course depends on the stimuli series, which may be different for each fMRI session even for the same experiment. Instead, we compare the z-maps. If the two components are for the same cognitive tasks, then the same functional regions ought to be activated, thus acquire relatively large z-values.

We then select the "most important" voxels. Intuitively, we should select voxels whose z-values are "large enough". In fact, here we always consider the absolute values, because in ICA, a negative sign could be assigned to both the time course and the z-map. Voxels with low absolute z-values are more likely to be random than causally related to the stimulus. Of course, time and space cost increases as more voxels are included. There is no consensus on what z-value should be considered "large enough", so here we simply take top 1% of the voxels with largest absolute z-values. Although this decision seems arbitrary, it has the virtue that the number of voxels representing different datasets tends to be nearly the same, which would not be true if a threshold were set based on the z-value itself. We call the selected voxels the *thresholded z-map*.

Although we select the 1% of the voxels to represent the t-map, there is still a potential artifact. In some experiments, only a part of the brain is scanned, and different experiments may have different parts scanned.

This will make retrieval results artificially good, since differences in the portion of the brain images become a surrogate for the underlying cognitive conditions. To address this problem, all feature selection and matching in this paper are conducted in the common part of all brains. That is, a chosen voxel has to be in the intersection region of all brains in database. The is in line with the method used in Mitchell et al. [18].

As a first step, we define the similarity between two components as the size of the overlap of their thresholded z-maps divided by the size of their union. That is, for two thresholded z-maps A and B, the similarity

$$similarity(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|}$$

We are also trying other metrics.

4 Testing and Results

In this preliminary research, we use data from 5 different experiments. The list of experiments and brief descriptions appears in Table 1.

4.1 Performance Evaluation

Our testing scheme is built on a standard information retrieval framework, in which every image is used as a query, and performance is evaluated by checking the returned ranked lists. A retrieved image is considered "relevant" to the query only if they are both for the same type of experiments.

As noted, different experiments have different numbers of datasets. In this case, average precision will not behave as one might like, and we are suspicious of using it. Instead, we use the "area under the ROC" [16] (for the sake of simplicity, we call this "ROC area") to evaluate a retrieval method. For a retrieved list, suppose the number of relevant elements is m and the number of non-relevant elements is n. The ROC curve starts at the origin $(0, 0)$. We traverse the ranked list from the top. If an element is relevant, the ROC curve goes up by a step $1/m$; otherwise, ROC curve goes to the right by a step $1/n$. If the area under the ROC is 0.5, then the retrieval method is no better than random selection. Generally, the retrieval performance is considered good if the area under the ROC is greater than 0.8. As noted, we use each of the datasets as a query against the rest, and report the average area under the ROC as the performance indicator.

To avoid the effect of intra-subject similarity, we exclude all the images of the *same subject* as query from the retrieved ranked list when we evaluate retrieval performance.

Table 2. Average ROC area for 360 datasets. 10 components are selected with LFC, HFC, and RDM, respectively. Two image similarity metrics, MAX and MWB, are included.

Method	MAX	MWB
Low Frequency	$.666 \pm .005$	$\mathbf{.729 \pm .005}$
High Frequency	$.645 \pm .006$	$.665 \pm .006$
Random Components	$.592 \pm .004$	$.716 \pm .006$

4.2 Results

Besides the method we just proposed, we also tested several other methods as comparisons. For component selection, we compare the performance of 10 "lowest expected frequency" components (LFC) with 10 "highest expected frequency" (HFC) and 10 "randomly chosen" components (RDM). For image matching, we include a method in which the similarity between dataset is the maximum of the component similarity, which is referred to as *MAX* (in Figure 2, the similarity between the two datasets is 0.6) as a comparison to the maximum weight bipartite matching (MWB). Table 2 shows the average area under the ROC. We can see that (1) the scores of MWB are always significantly better than MAX. (2) score of "low frequency components with MWB" is the highest, and it is significantly better than all others.

Figure 3 shows the average ROC area for individual experiments. We can see the LFC with MWB has the most stable performance across all the experiments. Note that RDM with MWB gets a high score in Table 2, but it is only because this method does well for the experiment with largest size (Morality, 248 datasets), so the average is pulled up, although it performs poorly for other experiments.

All 6 methods based on ICA do rather well on the largest category. But low frequency components, coupled with aggressive matching of each component to its best "partner" does the best, or very well on all categories of experiment. This is consistent with the facts noted earlier: (1) when we observe the JRF we expect low frequency to be most informative and (2) since the ordering of the components is not logically fixed, pairing each component to its best match for another experimental run gives a more complete picture of the similarity.

14

Table 1. Experiments

| Exp	Description	Papers	# of datasets
Oddball-auditory	Recognition of an out of place sound		4
Oddball-auditory	Recognition of an out of place sound		4
Event perception - House Active	Watching either a real film of a human being	[23]	28
Event perception - Study Active	Watching either a cartoon film	[23]	25
Morality	Deciding about problem situations having or lacking combinations of moral and emotional content	[12]	248
Study-Recall	Study and recall or recognition of faces, objects and locations	[19]	27
Recall-Only	Recall faces, objects and locations	[19]	9
Romantic	People in love see pictures of their important others, or of non-significant people	[1]	15
Total			360
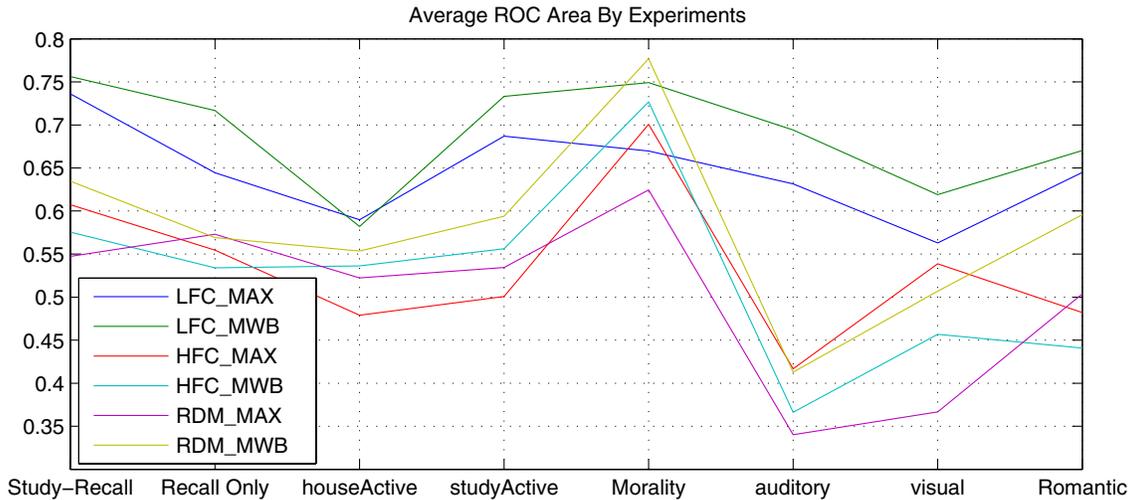
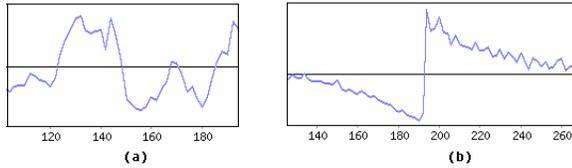**Figure 3. Average ROC area for separate experiments.**

15

Figure 4. Two time courses with low frequency spectrum. (a) is a hemodynamic response. (b) is head motion artifacts

5 Discussion and Future Work

Refining the "low frequency" heuristics In our experiment, the low frequency approach usually gives high rank to meaningful time courses, as shown in Figure 4 (a). However, in a few cases it also favor apparent bad time courses. Figure 4 (b) shows one such example. It is usually referred to as an artifact caused by head motion (note we have performed motion correction, but the effect still appears occasionally). The prior knowledge of these patterns maybe valuable assistance to low frequency component selection. Also, the number of components to select is also a question. The number 10 we used in this paper is one that gives good results. However, a fixed number for every experiment seems arbitrary and inflexible. We can improve the signal-noise ratio in features if the number task-related components can be identified for each dataset, with frequency and/or other heuristics.

6 Conclusion

We proposed a new retrieval algorithm for fMRI images, based on independent component analysis and a new similarity measure. In our preliminary study, the results show that (1) ICA can provide good features for fMRI image retrieval. (2) The component selection by low frequency spectrum of time courses, along with maximum weight bipartite feature matching, can be a powerful indicator for image similarities in the sense of cognitive tasks.

References

[1] A. Aron, H. Fisher, D. Mashek, G. Strong, H. Li, and L. Brown. Reward, motivation, and emotion systems associated with early-stage intense romantic love. *J Neurophysiol*, 94:327–337, 2005.

[2] B. Bai, P. Kantor, N. Cornea, and D. Silver. Ir principles for content-based indexing and retrieval of functional brain images. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM06)*, 2006.

[3] B. Bai, P. Kantor, N. Cornea, and D. Silver. Toward content-based indexing and retrieval of functional brain images. In *Proceedings of the (RIAO07)*, 2007.

[4] C. Beckmann and S. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2), 2004.

[5] J. Bondy and U. Murty. *Graph Theory with applications*. Macmillan, London, 1976.

[6] V. Calhoun, T. Adali, L. Hansen, J. Larsen, and J. Pekar. Ica of functional mri data: An overview. In *4th Int. Sym. on Indep. Comp. Analy. and Blind Signal Sep*, 2003.

[7] Y. Cheng, V. Wu, R. Collins, A. Hanson, and E. Riseman. Maximum-weight bipartite matching technique and its application in image feature matching. In *Proc. SPIE Visual Comm. and Image Processing, Orlando, FL*, 1996.

[8] J. Ford, H. Farid, F. Makedon, L. Flashman, T. McAllister, V. Megalooikonomou, and A. Saykin. Patient classification of fmri activation maps. In *6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention*, 2003.

[9] J. Fox. *Linear statistical models and related methods*. 1984.

[10] R. Frackowiak, K. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. Penny. *Human Brain Function (2nd Edition)*. Elsevier Academic Press, 2004.

[11] K. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1994.

[12] J. Greene, R. Sommerville, L. Nystrom, J. Darley, and J. Cohen. An fmri investigation of emotional engagement in moral judgment. *Science*, 293, 2001.

[13] D. Hu, L. Yan, Y. Liu, Z. Zhou, K. F. KJ, C. T. C, and D. Wu. Unified spm-ica for fmri analysis. *Neuroimage*, (3):746–55, 2005.

[14] A. Hyvarinen, J. Karkunen, and E. Oja. *Independent Compoenent Analysis.* John Wiley and Sons, 2001.

[15] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu. Support vector machines for temporal classification of block design fmri data. *NeuroImage*, 26:317–329, 2005.

[16] S. Mason and N. Graham. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc*, 30:291–303, 1982.

[17] M. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. Kindermann, T.-W. Lee, and T. Sejnowski. Analysis of fmri data by decomposition into independent spatial components. *Human Brain Mapping*, 6:1–31, 1998.

[18] T. Mitchell, R. N. R. Hutchinson, F. Pereira, and X. Wang. Learning to decode cognitive states from brain images. *Machine Learning*, 57:145–175, 2004.

[19] S. Polyn, J. Cohen, and K. Norman. Detecting distributed patterns in an fmri study of free recall. In *Society for Neuroscience conference*, San Diego, CA, 2004.

[20] S. Smith, P. Bannister, C. Beckmann, M. Brady, S. Clare, D. Flitney, P. Hansen, M. Jenkinson, D. Leibovici, B. Ripley, M. Woolrich, , and Y. Zhang. Fsl: New tools for functional and structural brain image analysis. In *Seventh Int. Conf. on Functional Mapping of the Human Brain. NeuroImage*, volume 13, 2001.

[21] K. Thulborn, C. Martin, and J. Voyvodic. fmri using a visually guided saccade paradigm in alzheimer's disease. *AJNR Am J Neuroradiol*, 21:524–531, 2000.

[22] M. Tipping and C. Bishop. Probabilistic principal component analysis, 1997.

[23] A. Zaimi, C. Hanson, and S. Hanson. Event perception of schema-rich and schema-poor video sequences during fmri scanning: Top down versus bottom up processing. In *In Proceedings of the Annual Meeting of the Cognitive Neuroscience Society*, 2004.

[24] J. Zhang and V. Megalooikonomou. An effective and efficient technique for searching for similar brain activation patterns. In *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2007.