# Toward Content-based Indexing and Retrieval of Brain Images

Bing Bai, Paul Kantor
bbai, kantorp@cs.rutgers.edu
Dept. of Computer Science
Rutgers University

Nicu Cornea, Deborah Silver
cornea, silver@caip.rutgers.edu
Dept. of Electrical and Computer Engineering
Rutgers University

## Abstract

In this paper, we explore the concept of a "library of brain images", which implies not only a repository of brain images, but also efficient search and retrieval mechanisms that are based on models derived from IR practice. As a preliminary study, we have worked with a collection of functional MRI brain images assembled in the study of several distinct cognitive tasks. We adapt several classical IR methods (inverted indexing, TFIDF and Latent Semantic Indexing(LSI)) to content-based brain image retrieval. Our results show that efficient and accurate retrieval of brain images is possible, and that representations motivated by the IR perspective are somewhat *more* effective than are methods based on retaining the full image information.

## Introduction

fMRI (functional Magnetic Resonance Imaging) [18, 11, 6] is a technique used to "monitor" brain activities. The most widely used fMRI method is BOLD (blood oxygen level dependent) fMRI. In this technique, 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 brighten the brain region in the image (as shown in Figure 1). 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). We call the former status the *condition*, and the latter status the *control*. A 3D brain scan is made every few seconds, during each experimental "run" 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.

As a method for watching "how the brain works", fMRI has been used as a powerful research tool in many aspects of neuroscience studies in the past decade. More recently, fMRI

Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007 - Copyright C.I.D. Paris, France

↑ neural activity ➔ ↑ blood oxygen ➔ ↑ fMRI signal

Figure 1: Neural activities cause stronger (brighter) fMRI signals

is gaining clinical attention. For example, studies [24] 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. More importantly, 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.

With this idea of sharing in mind, an online fMRI library – the Dartmouth fMRI data center (fMRIDC) [25] – was founded in 2003. As of May 2006, the fMRIDC has archived brain images from 120 experiments and has received 2000 data requests. This valuable data is presently indexed with textual descriptions (also called "textual meta-data"), including the technical configurations under which the images were collected, subject information, information about the cognitive tasks (provided via related publications) and any other information the providers consider important.

In this paper, we report some investigations of *content-based* indexing of fMRI images, which is validated by using *condition* 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:

- Helping researchers find similar studies and related research work.
- Helping researchers discover hidden similarities among superficially different studies.
- Helping doctors diagnose brain disorders, by looking at the clinical history of persons with similar fMRI patterns.

One straightforward idea is to map the brain image retrieval problem to the framework of content-based image or video retrieval. However, common features used in 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. Moreover, there is no human gold standard for fMRI image classification. 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.

We describe here an indexing scheme based on selection of voxels, which are described as *activated* regions. This substitution of the regions of the brain, to "stand for" the cognitive processes is somewhat analogous to the use of “terms” to stand for “concepts” in the field of textual information retrieval. In fact, while one can “talk about” a specific concept without using the common words associated to that concept, under current thinking it is most unlikely that a given (normal, healthy) brain would perform a basic cognitive task without using certain associated regions.

With this correspondence in mind, we show that using the classical information retrieval techniques of inverted indexing (applied not to terms, but to “activated voxels”), retrieval by condition can be implemented effectively and efficiently. The inverted index provides the infrastructure for many different Information Retrieval algorithms. In this note we examine two of the most basic, TFIDF similarity [21] and LSI representation and similarity [3].

Our contributions in this paper can be summed up as follows:

- To the best of our knowledge, this is the first time that 4D (the 4th dimension is time) fMRI images have been put into a content-based retrieval framework, although there is a lot of closely related work (on retrieval of 3D medical images [16] and on brain image classification [15, 4, 12]).
- We have adapted several textual retrieval methods (inverted indexing, TFIDF and LSI) to fMRI images, and tested them on a database of moderate size and considerable variety.
- We have introduced “fuzzy indexing” into fMRI image indexing/retrieval to control the effect of inter-subject structural inconsistency, and find that it yields the best performance among the tested methods.

We also describe, briefly, the processing pipeline and the testing framework.

## Background

### Analysis of functional MRI images

The most widely used mathematical method for fMRI research is the general linear model (GLM) [7, 5]. In the GLM, the time series of intensity 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 called hemodynamic response function (HRF). In the work reported here the default HRF of the fMRI software package FSL[23], which is a mixture of two gamma-distribution shapes is used throughout.

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 we find 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. The entire regression analysis is performed for each voxel, and a brain map which shows, at each voxel, the corresponding t-value is called the t-map for the stimulus. As shown in Figure 2, a brighter voxel represents a larger t-value, and thus a more significant response to the stimulus.

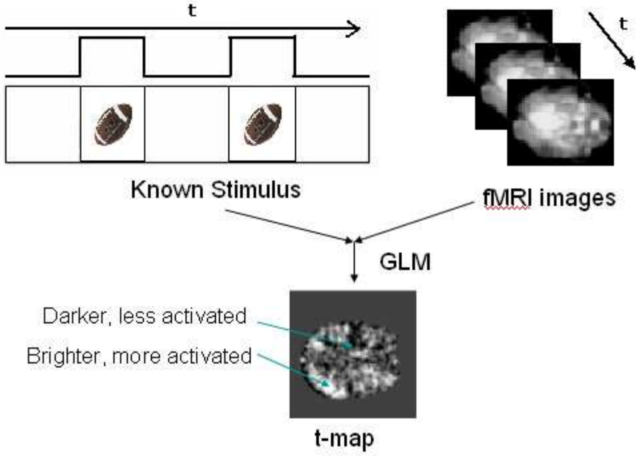

Figure 2: A t-map is built by applying the GLM to the brain image and the known stimulus. The value of a voxel in the t-map indicates how significantly this voxel responds to this stimulus.

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) [10, 14]. Like GLM, ICA models fMRI signals in terms of underlying or "latent" components. However, ICA does not assume the knowledge of either the stimulus or the hemodynamic response function. Instead, ICA postulates that there are "a small number of" statistically independent latent components, and finds the linear transformation (from voxels to components) that maximizes the independence between components. Note that with this approach (as with the corresponding Factor Analysis used in many studies), it may be hard to say, of the underlying ICA components, whether or how they are related to the cognitive tasks.

### Classification of Brain Images

Some other related work is presented as "classification" of brain images. Some researchers try to detect "activation" volumes in the same brain image sequence [15, 12], while others try to distinguish experiments with different conditions [4].

In these studies, machine learning (ML) methods (k-nearest neighbors [15], Bayesian [15], Support Vector Machines (SVM) [15, 12] and Fisher's Discriminant Analysis [4]) are applied using features such as the time series of voxels [15, 12], or to intermediate result from other processing, such as t-maps [4]. All ML methods characterize the distribution of some set of features for labeled training datasets, and use these characteristics to classify other datasets.

In the present work, we recognize that classification must involve the same underlying problems (representation and similarity) as does retrieval. But, in addition to considering the accuracy of classification, we consider retrieval efficiency and scalability to large databases.

### Information Retrieval Techniques

Classical information retrieval approaches were developed for text retrieval. The most fundamental technical innovation is inverted indexing [26]. Inverted indexing supports fast retrieval based on query terms. Considering that terms that occur very frequently may be of little value in discriminating among texts, TFIDF [21] was designed to give terms weights by both local (term frequency in a document) and global importance (inverse to the number of documents in which the term occurs, denoted as “IDF”, for inverse document frequency). Latent Semantic Indexing [3] reduces the dimensionality of the term-document matrix by finding principal components, here called “latent semantic components”. These ideas have in fact been applied to image retrieval problems by appropriately mapping the visual features to terms [20, 9]. One may say that their applications in this note defines “latent cognitive activities”, realized in patterns of brain activations.

## Data Processing and Methods

### Term Definitions

In fMRI, there is considerable variation in the use of terminology and we briefly define the use of some key terms in this note. Note that these definitions may not be consistent with those in all other papers.

1. *Run*: A 3D brain image series that is scanned during one experimental session, with one subject continuously under observation.
2. *Condition*: A chosen stimulus. A typical run will contain at least two different conditions, one of which may be designed as a control.
3. *Experiment*: An fMRI study undertaken for cognitive research. An experiment will generally have many runs, with multiple subjects.
4. *Dataset*: A t-map built by selecting, from a single run, the regions that respond differently to one specific condition, versus all the other conditions (including control).

Because there can be several conditions in the same run, that is, the subject could conduct different tasks in the same run, several datasets may be generated from the same run.

### Preprocessing and Registration

The raw fMRI images from different scanners, or for different subjects are not comparable. Raw signals contain many kinds of noise. Thus fMRI images go through the following processing steps before either GLM or ICA is applied:

1. Apply motion correction to align all 3d images in a time sequence 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. Note that registration is not necessarily a preprocessing step. In this work, we choose to do registration after the activation maps are generated. 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 all these steps using the fMRI software package FSL [23].

### Feature Selection

The features we seek, which will play the role of "terms" in retrieval should satisfy two criteria:

1. *Effectiveness*: Describe the activated regions for certain conditions.

2. *Efficiency*: Can be embedded into efficient algorithms.

We propose a feature selection method, as shown in Figure 3. This scheme can be viewed as a two-stage black box system.

The first stage is to build "activation maps", in which the value of each voxel indicates the level of activation of the voxel. We currently use t-maps generated by the GLM module of FSL, but this could be replaced by other activation maps. Building activation maps is at the heart of much fMRI research and many methods are still under development. In fact, it is possible that the best methods for different experiments will be different. However, for the simplicity of discussion, we will use only the "t-map" in following.

The second stage selects the "most important" voxels. Intuitively, we should select voxels whose t-values are "large enough", for voxels with low t-values are more likely to be random than causally related to the stimulus. Of course, time and space cost increase as more voxels are included. There is no consensus on what t-value should be considered "large enough",

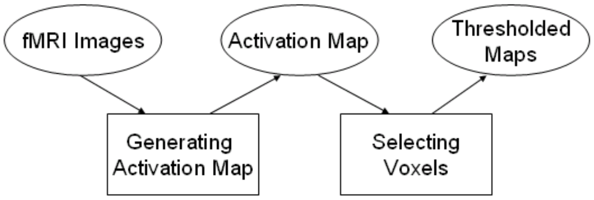

Figure 3: Feature selection is in two stages, the first stage is to build a brain map in which the value of each voxel indicates the activation, the second stage is to select most important of these voxels based on their activation.

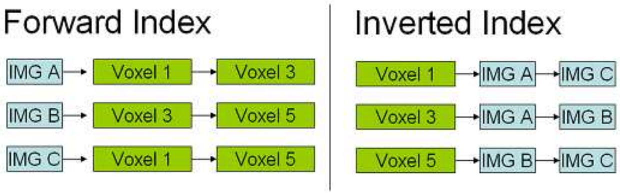

Figure 4: Forward and inverted index. Two voxels are activated in each brain images.

so here we simply take that 1% of the voxels with largest t-values. Although this decision seems arbitrary, it has the virtue that the number of voxel representing different datasets tends to be nearly the same, which would not be true if a threshold were set based on the t-value itself. We call the image of the selected voxels the *thresholded t-map*.

Our retrieval model closely parallels classical text retrieval, with these correspondences:

| | | |
|---|---|---|
| Brain images | *correspond to* | Documents |
| Activated voxels | *correspond to* | Terms |
| Query voxels | *correspond to* | Query Terms |

With this similarity in mind, we build both forward and inverted indices. In the forward index, each entry is an image, with a link to its activated voxels. In the inverted index, each entry is a unique voxel (in the standard brain), pointing to a list of brain images in which that voxel is activated. This is shown in Figure 4.

Although we select the top 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. Clearly this "trick" is beneficial for validation but can not scale indefinitely, as the intersection would eventually vanish.

### Similarity Measures

*Overlap*
The method we call "overlap" corresponds to "coordinate matching" [26]. In this method, the retrieved datasets are ranked according to the size of their overlap with the activated voxels in the query image. To maintain the analogy to "terms" in text retrieval, we refer to these as the "query voxels". For instance, a candidate dataset whose representation contains 50 of the query voxels will outrank a dataset that contains only 49 of query voxels.

Since activation is here binarized, the overlap method is also a variation of the cosine measure for vector similarities. The thresholded t-maps can be considered as sparse vectors, with value 1 for activated voxels, and 0 for the rest. It can be shown that if the number of "1"s (that is, the number of selected voxels) is the same for vectors $A, B, C$ and $D$, then

$$cos(A, B) > cos(C, D) \Leftrightarrow overlap(A, B) > overlap(C, D)$$

Thus, for the binarized representation with exactly the same number of active voxels, overlap and the cosine measure are equivalent in the sense of retrieved ranked lists.

*Fuzzy Overlap*
Now consider the following case. A specific voxel $r$ is active in two different brain images $A$ and $B$. However, the transformed voxels $r_a$ and $r_b$ do not overlap in standard space, due to differences in the shape of the brains, although they might be close. If we only index the datasets with the exact activated voxel (in this case, $A$ is in $r_a$'s entry, $B$ is in $r_b$'s entry), then retrieval system will miss $B$ if the query is $A$. To address this problem, we index a dataset $A$ not only with its activated voxel $r_a$, but also with the *neighbors* of $r_a$. Specifically, we say that a voxel $u = (x_1, y_1, z_1)$ is a neighbor of voxel $v = (x_2, y_2, z_2)$ if the L-infinity-norm $(L_\infty)$ $|u - v|_\infty < R_f$. $R_f$ is called the *fuzziness radius*. If $R_f$ is set too small, some related datasets may be missed. If $R_f$ is set too large, false alarms increase, and precision will drop. Figure 5 shows a 2D case with fuzziness radius 1. The voxel marked 8 is active in image A, so A is put in the index entries of voxel 8 and of its neighbors.

*Latent Semantic Indexing*
*Latent Semantic Indexing* (LSI) [3] is a technique to automatically discover similarities between terms and documents, and to apply this information to information retrieval. Suppose for $n$ documents, the total number of terms is $m \geq n$, we first build an $m \times n$ term-document matrix $M$. In this matrix, the element at location $(i, j)$ is the number of occurrences of term $i$ in document $j$. Singular value decomposition (SVD) [17] is applied to this matrix, and the following decomposition is generated:

$$M = U \cdot S \cdot V^T \tag{1}$$

$U$, a matrix of dimensions $m \times n$, is called the *term matrix.* The $n$ columns of $U$ are orthonormal. Each column is interpreted as a "latent semantic component"; each element in the column is the weight of the corresponding term in this component. On the other hand, $V$ is called the document matrix. The $i$th column of $V$ contains the weights of the "latent semantic components" in document $i$. The diagonal matrix $S$ contains the singular values, sorted in descending order.

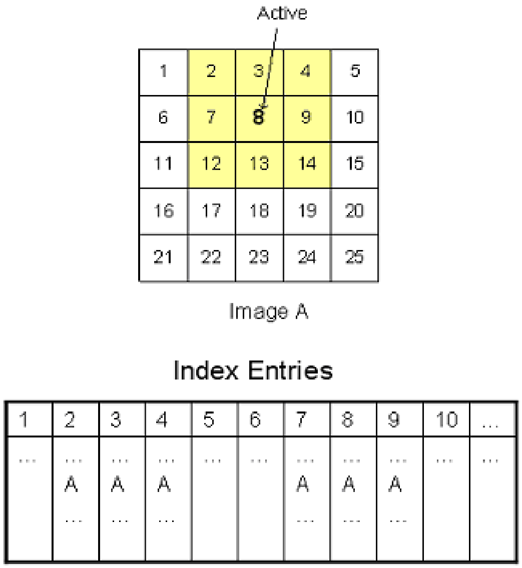

Figure 5: One example of fuzziness radius 1. Voxel 8 in Image A is active, A is indexed with voxel 8 and its neighbors

In the case of text, when a query is received, a term vector $q$ for the query is built first; each element in this vector is the number of occurrence of the corresponding term. Then this term vector is transformed into semantic space by projection onto the semantic components:

$$\hat{q} = q^T U S^{-1} \tag{2}$$

. The similarity between the query and the $i$th document is defined as the cosine of $\hat{q}$ and $V_i$.

$$\cos(\hat{q}, V_i) = \frac{\sum_j (\hat{q}_j \cdot V_{ij})}{\sqrt{\sum_j \hat{q}_j^2 \sum_j V_{ij}^2}} \tag{3}$$

As noted, 1 percent of the voxels, those with largest t-values are selected to be "activated". If a voxel is activated, the corresponding value in the term vector is set to 1, otherwise it is set to 0. The standard routine of LSI (equation 1, 2, 3) is applied after that.

*TFIDF*
Under TFIDF [21] the weight of a term is the product of two parts: TF(term frequency) and IDF (inverse document frequency). For each term $t$, TF indicates the importance of $t$ in a given document $d$, while IDF indicates the general importance of $t$. TFIDF has many

variations, and a popular form is the following:

$$tf(t,d) = \frac{n(t,d)}{\sum_k n(k,d)} \tag{4}$$

$$idf(t) = \log(\frac{D}{T}) \tag{5}$$

$$tfidf(t,d) = tf(t,d)idf(t) \tag{6}$$

$n(t,d)$ is the number of times term $t$ occurs in document $d$. $tf(t,d)$ is $n(t,d)$ divided by the document length. $D$ is the total number of documents in the collection. $T$ is the number of documents containing the term $t$.

We have to translate this to brain images. Unlike a term in a document, an activated voxel appears only once in any brain image. So if we mimic the behavior of term frequency, the voxel/term frequency is always 1. To give a finer representation, we can replace the term frequency with the t-value to indicate the local importance of the term.

IDF translates naturally to fMRI images. If a voxel is activated in too many images, it is obvious that this voxel is not a discriminant feature. One source of such voxels are vessels, especially arteries. Oxygenated blood flow has to pass through these vessels to reach the actual activated regions.

Once the weights for all the voxels are known, a sparse vector is built for every thresholded t-map; the t-value for each voxel is its weight. Again the similarity measure between two thresholded t-maps is defined as the cosine of the vectors.

## Testing and Results

In this section, we introduce the evaluation scheme, the experimental database and the experimental results.

### Datasets

In this preliminary research, we use 430 datasets from 5 real fMRI experiments. fMRI data is relatively harder to acquire than textual documents, but we managed to make the size of the database reasonable for image retrieval task. The list of experiments and brief descriptions are shown in Table 1. The conditions of these experiments are shown in Table 2.

In some of the experiments, namely, "Event perception" and "Oddball", different conditions are performed only in separate runs. For the rest, several conditions will occur in each run. We note that the retrieval result could be affected by factors representing "same run", and "same subject", as well as "same condition", which is what we want. 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.

### Performance Evaluation

| Exp | Description | Publication |
|---|---|---|
| Oddball | Recognition of an out of place image or sound | |
| Event perception | Watching either a cartoon movie or real film of a human being | [27] |
| Morality | Subjects make decisions about problem situations having or lacking combinations of moral and emotional content | [8] |
| Recall | Study and recall or recognition of faces, objects and locations | [19] |
| Romantic | People in love see pictures of their important others, or of non-significant people | [1] |

Table 1: Experiments

Our testing scheme is built on an information retrieval framework. We use every image as query, and evaluate the performance of our method by checking the returned ranked lists. A retrieved image is considered "relevant" to the query only if they are both for the same experimental condition. As noted, different experiments have different numbers of datasets. In this case, average precision will be sensitive to the data size of each condition, and make the comparisons among conditions difficult (more details can be found in discussion). Thus, we use the "area under the ROC" [13] (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 (excluding same subject), and report the average area under the ROC as the performance indicator.

### Results

In Table 3 we list the average ROC area for 5 different similarity measures. We regard "cosine" as the baseline system. "Cosine" takes whole t-maps as big vectors where each voxel is an element in these vectors; the similarity between two t-maps is simply the cosine of the angle between the vectors.

Except for the baseline system, all methods select only 1 percent of the voxels to be labeled activated. Using the Bonferroni correction for multiple comparisons, almost all of the pairwise differences in Table 3 are significant at the 95% confidence level. With 5 methods there are 10 comparisons, and the corrected 95% confidence interval is 2.57 times the standard error of the mean. Thus differences of mean ROC areas that are at least $2.57 * 0.007 = 0.018$ are statistically significant at the 95% confidence level. Eight of the ten contrasts are significant at or above this level, and the difference between F2 and LSI is very very close

| **Experiment** | **Condition** | TR(s) | Volumes | **# of datasets** |
|---|---|---|---|---|
| Oddball | Auditory | 2.0 | 150 | 4 |
| | Visual | 2.0 | 150 | 4 |
| Event Perception | House Active | 1.5 | 110 | 28 |
| | Study Active | 1.5 | 210 | 25 |
| Recall | Study Face | 1.8 | 510 | 27 |
| | Study Object | 1.8 | 510 | 27 |
| | Study Location | 1.8 | 510 | 27 |
| | Try to think of Face | 1.8 | 510 | 27 |
| | Try to think of Object | 1.8 | 510 | 27 |
| | Try to think of Locaction | 1.8 | 510 | 27 |
| | Recall Face | 1.8 | 103 | 9 |
| | Recall Obj | 1.8 | 103 | 9 |
| | Recall Loc | 1.8 | 103 | 9 |
| Morality | M+E+ | 2.0 | 150 | 50 |
| | M+E- | 2.0 | 150 | 50 |
| | M-e- | 2.0 | 150 | 50 |
| Romantic | Neutral Face | 5.0 | 144 | 15 |
| | Positive Face | 5.0 | 144 | 15 |
| **Total(N)** | | | | 430 |

Table 2: Datasets and Numbers

| **Method** | **Mean ROC area** | **Standard deviation** | **Standard Error of the Mean** |
|---|---|---|---|
| Cosine | .704 | .148 | .007 |
| Overlap (Fuzziness Radius 0) | .733 | .152 | .007 |
| Overlap (Fuzziness Radius 2) | **.772** | .135 | .007 |
| TFIDF | .735 | .149 | .007 |
| LSI(10 components) | .755 | .134 | .006 |

Table 3: Average ROC area for 430 datasets.

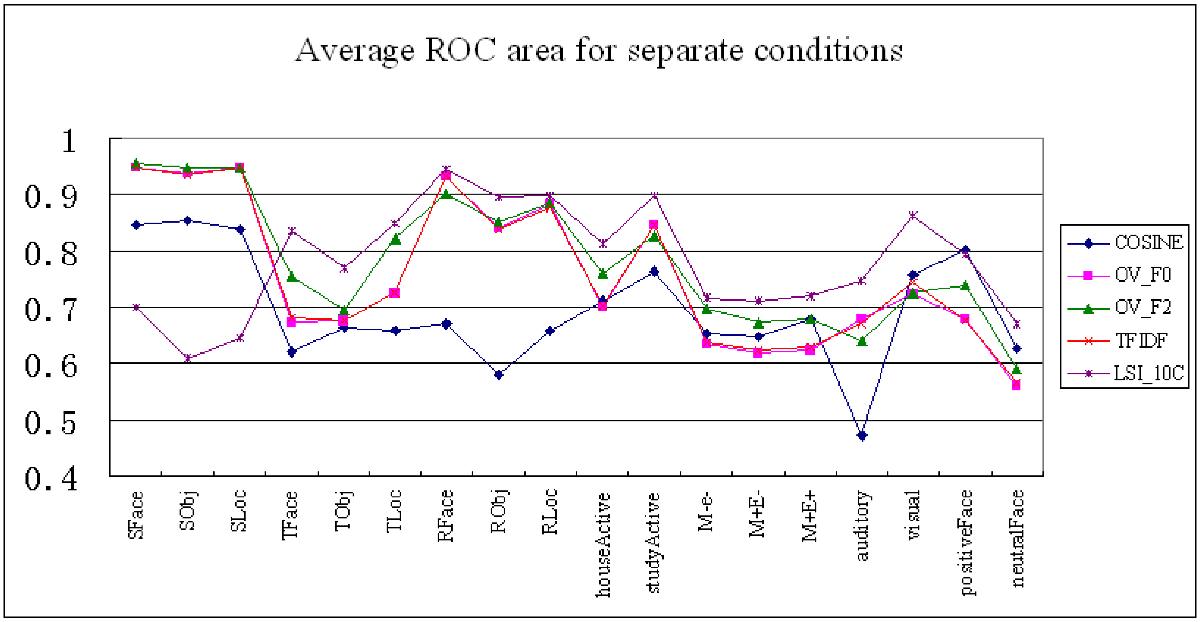

Figure 6: Average ROC area by separate conditions. OV_F0 is simple overlap (radius 0). OV_F2 is fuzzy overlap with radius 2.

(the difference is 0.017). The exception is the pair TFIDF and Overlap (Radius 0). This lack of difference suggests that perhaps moving from binarized voxels to weighted voxels will not be an important factor in retrieval.

We see that all methods based on thresholded "t-maps" perform better than the baseline method significantly. That provides support for the argument that an threshold for t-values is necessary. As the threshold is lowered, we are more likely to be including random noise, which will hurt the quality of retrieval, rather than improving it. More importantly, this thresholding approach makes inverted indexing quite efficient. We will conduct further study on threshold selection in the future.

We can see from Table 3 that overlap with fuzziness radius 2 has the best performance. Although we do not list the results for radius 1, 3 and 4 here, the relationship between these results is $R(0) < R(1) < R(2) > R(3) > R(4)$, where $(R(i)$ means the average ROC area with radius $= i$). This is qualitatively as we expected. An appropriate amount of fuzziness can correct for some registration error, and for anatomical variations among individual brains, while too much fuzziness eliminates the difference between the activated and non-activated regions.

The performance of LSI is known to depend on the number of components retained in indexing. In this study we find that the average ROC area reaches its maximum when we consider only the first 10 components. As shown, while this is better than simple overlap (fuzziness radius 0), it is not as good as the fuzzy overlap of radius 2.

Figure 6 shows average ROC area for different conditions separately. Due to their close relations to the "cosine measure", the 3 methods (Overlap/Fuzzy Overlap/TFIDF) have similar performance for each condition. But the profile of results for LSI is quite different. It is interesting that the 3 conditions exhibiting the best performances with cosine-related methods (SFace, SLoc, SObj, which means: watching human faces, locations and objects, respectively), have relatively low scores in LSI. On the other hand, LSI performs generally better for other conditions. We are still trying to understand this fact.

### Algorithm Effiency

*Space Cost*

For Simple overlap, suppose the number of selected voxels in each dataset is $S$, the number of all voxels (in our case it is the intersection of all images) is $W$. The inverted indexing method reduces the space by a factor $2S : W$. In our case, $S/W$ is 0.01 since we take only 1% of the voxels. There is a factor of 2 because we need both forward and inverted indices. In a forward index, each element is a voxel ID. Similarly, each element in an inverted index is a image ID. Both of them can be represented by a 4-byte integer.

For the fuzzy overlap method, the number of indexed voxels will increase because neighbors are counted. If the neighbors did not overlap, the number of indexed voxels for fuzzy overlap with radius 2 would be 125 times ($5^3$) as many as that of simple overlap. In fact, activated voxels are often clustered, and thus many of neighbors actually overlap. In our experiment, as an empirical example, the space ratio of fuzzy overlap with radius 2 to simple overlap is 11:1.

For TFIDF, in addition to the voxel IDs and document IDs, we also need to store t-values for each indexed element. This doubles the space because each t-value is a 4-byte floating point number.

For LSI, we do not use the inverted indexing. Instead, the SVD triplet $(U, S, V)$ is stored. Suppose the number of voxels is $m$, the number of datasets is $n$, and we keep the top $t$ components, then the space needed is $m \times t + n \times t + t$. In our experiment $t$ is only 10. Note that $m$ is not the number of voxels in the whole image, it is the number of voxels that are activated in at least one image. In fact, these two numbers are very close in our experiment.

*Time Cost*

First, let us check the time cost of the quasi-cosine measures, including simple/fuzzy overlap and TFIDF. Let the number of selected voxels for each dataset be $S$, the number of images is $n$, and the number of voxels in the brain space is $V$. Then the expected length for each index entry is $L = NS/V$. When a query is submitted (with $S$ voxels, of course), we need to traverse $S$ lists in the inverted index to generate the ranked list. Each traversal takes $O(L)$ time. So in the average case, one single retrieval takes time:

$$O(SL) = O(S^2N/V) \tag{7}$$

To see how efficient this algorithm is, we compare it with another overlap-counting algorithm. We call this a "merge method" [22], for reasons which become clear below. In this algorithm, the overlap is calculated as following:

- Step 1: Convert the 3D coordinates (x,y,z) of "important voxels" to " 1D coordinates" with the equation $I(x, y, z) = zXY + yX + x$, in which $X, Y, Z$ are the width, length and height of a brain volume.
- Step 2: Sort the "important voxels" by the "1D coordinates", let us call the result "feature lists".
- Step 3: When we compare two data sets A and B, we take the feature lists of A and B, and traverse the two list from the top in a manner similar to a merge sort, and count the overlap on the way.

We see that steps 1 and 2 can be considered as "preprocessing". Once this preprocessing is done, a matching between two feature lists takes $O(S)$. To generate a ranked list for a single retrieval, $N - 1$ such comparisons have to be performed, thus the time cost is $O(SN)$. In our case, $S/V$ is 0.01, thus equation 7 can also be written as $O(SN)$, with a possible smaller coefficient.

We compared these 2 methods with a task which contains 2560 retrieval tasks in a database of 430 datasets. The "merge" method took 20 minutes to do preprocessing, and 3 minutes to do retrieval, while the inverted indexing method took only 20 seconds to index, and another 20 seconds to do retrieval. "Preprocessing" or "indexing" can be considered as one time costs, and we can afford either way even with a much larger collection. But the ratio *(1:9)* between retrieval times is really impressive.

Finally, let us look at the time cost of indexing and retrieval in LSI. Indexing involves a singular value decomposition of a large voxel-image matrix. This may take some time, but it is a one-time cost. Also, we can extract only a number of leading components, instead of conducting a full SVD. We use the same notations as in the previous section, i.e., $m$ is the number of voxels, $n$ is the number of images, and $t$ is the number of singular values kept. The cost for a single retrieval is the time to represent the query in the component space, which is $t \times m$ multiplications, plus the time to calculate the cosine for that representation with all image vectors, which is $t \times n$. So the total time for a single retrieval is $t(m + n)$ multiplications.

## Discussion and Future Work

### Measure of Precision in Brain Image Retrieval

These results suggest that approaching the analysis of brain images from the perspective of indexing and retrieval may result in strong improvements in the efficiency of processing and comparison, with no loss in the effectiveness of the representations. In fact, if we accept the notion of "same condition" as a gold standard, these methods, particularly fuzzy overlap with radius 2, have resulted in an improvement in effectiveness, as measured by the mean area under the ROC.

We consider briefly the relation of these results to the more conventional IR measures such as precision at the 10-th retrieved image ($P_{10}$). Figure 7 shows that using $P_{10}$ will obscure some of the striking differences noted in Figure 6. In addition, we have hesitated to present results in terms of p-values because of the dependence of those values on the composition of our test collection.

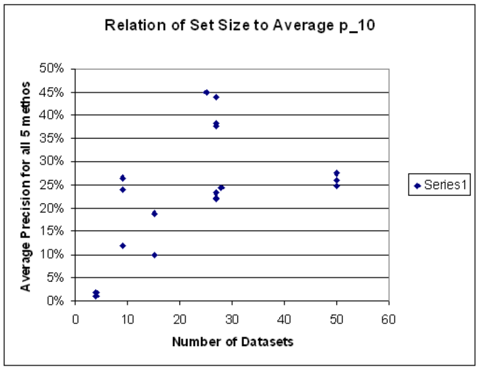

Figure 7: Relation of average $P_{10}$ and dataset size for each condition.

This fact is illustrated by Figure 7 and Table 4. This plot shows that there is a general trend for $P_{10}$ to increase with the number of cases corresponding to the condition. On this plot we see that there are two conditions with extremely high $P_{10}$ (SFace and studyActive), two with extremely high ratios of $p_{10}/N_{cases}$ (RFace and RLoc) and five with extremely low values of this ratio (auditory oddball, visual oddball, M+E-, M+E+, M-e-). The two "oddball" experiments have small datasets. The Morality data present some difficult challenges which will be discussed more fully elsewhere. In sum, we feel that the plot of $P_{10}$ versus the size of the dataset may prove useful, but that $P_{10}$ alone does not accurately summarize the effectiveness of an indexing and retrieval system for brain images.

### The Meaning of LSI in Brain Images

In section , we motivated LSI simply by an analogy between textual retrieval and brain image retrieval. Nevertheless, we suspect that LSI has important anatomical and cognitive meanings that might lead to better ways to index.

First, it is quite possible that a cognitive neural process is decomposed into sub-tasks by the brain. For example, many stimuli are delivered as visual signals. So the corresponding brain activations should include the basic visual reaction, recognition of the objects, and/or higher level cognitive processes. Second, oxygen is conveyed by vessels, and big vessels like arteries will be activated for many different stimuli. With LSI, we may hope to isolate these factors, and link the LSI basis vectors to more fundamental cognitive processes, while providing retrieval for a wide range of queries.

This may lead to consideration of other factor analysis methods such as PCA [2]. In fMRI, PCA was used to reduce dimensionality in [4]. In face recognition applications, PCA is used

| **Condition** | **N** | **AP10** | **AP10/N** |
|---|---|---|---|
| Auditory | 4 | 1% | 0.25 |
| Visual | 4 | 2% | 0.50 |
| House Active | 28 | 24% | 0.87 |
| Study Active | 25 | **45%** | 1.80 |
| SFace | 27 | **44%** | 1.62 |
| SObj | 27 | 38% | 1.40 |
| SLoc | 27 | 38% | 1.42 |
| TFace | 27 | 22% | 0.83 |
| TObj | 27 | 22% | 0.82 |
| TLoc | 27 | 23% | 0.87 |
| RFace | 9 | 26% | **2.94** |
| RObj | 9 | 12% | 1.31 |
| RLoc | 9 | 24% | **2.67** |
| M+E+ | 50 | 27% | 0.55 |
| M+E- | 50 | 26% | 0.52 |
| M-e- | 50 | 25% | 0.50 |
| Neutral Face | 15 | 10% | 0.66 |
| Positive Face | 15 | 19% | 1.25 |
| **Total** | 430 | | |

Table 4: Average $P_{10}$ for all conditions. N is the number of datasets. AP10 is the average $P_{10}$ of the 5 matching methods.

to reduce the dimensionality by building components called “eigen-faces”. In the similar fashion, we can perhaps build “eigen-t-maps”.

### Pair-wise Intersection

In the analysis described here, we use the intersection of all the datasets. With more and more experiments included, this intersection will become smaller and smaller, finally becoming empty.

To address this problem, we propose using “pair-wise intersection”. That is, the similarity between two datasets depends on the intersection of these two datasets, not their intersection with others. Note that this makes the inverted indexing somewhat more complicated, as “enough components” must be retained to ensure that the intersection will be of the desired size (1% of the brain).

### Inverted Retrieval

Another interesting direction is what we may call “inverted retrieval”. Instead of retrieving “similar images”, we can retrieve “similar voxels”. This is also potentially valuable in fMRI research. If we have already identified the functionality of a brain region, we might want to seek other regions with the same or similar function.

In the IR framework, this task can be naturally implemented. For overlap/TFIDF approaches, we simply switch the roles of forward and inverted indices. For a given“query voxel”, we rank the voxels in decreasing order of the number of images in which they were “activated” along with the query voxel. This same notion can also be extended to LSI by interchanging the “term/voxel matrix” $U$ and the “document/image” matrix $V$. Quite generally, in retrieval algorithms like inverted indexing or LSI, “terms/voxels” and “documents/images” have a “dual” relationship, and an application in one direction can often find a meaningful counterpart in the other direction.

## Conclusion

In this preliminary study, we show that efficient fMRI retrieval can be implemented by adapting classical textual retrieval techniques and using inverted indexing. The results show that retrieval by condition with high precision is possible. Using thresholded t-maps and, in particular, using what we have loosely called fuzzy indexing helps to improve retrieval performance.

## Acknowledgments

This work is supported by NSF grant ITR 0205178. The authors are grateful to Ali Shokoufandeh and Sven Dickinson for their helpful advice and encouragement.

# 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] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] 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.

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

[6] 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.

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

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

[9] J. Hare and P. Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In *Proceedings of 4th International Conference on Image and Video Retrieval.* Springer, 2005.

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

[11] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, and R. T. et al. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. In *Proc. Nat. Academy of Sciences USA*, volume 89, pages 5675–5679, 1992.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. a review of content-based image retrieval systems in medicine - clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23, 2004.

[17] J. Nash. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, 2nd ed.* Adam Hilger, Bristol, England, 1990.

[18] S. Ogawa, D. Tank, R. Menon, J. Ellermann, S. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proc. Nat. Academy of Sciences USA*, 89:51–55, 1992.

[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] P. Praks, J. Dvorsky, and V. Snasel. Latent semantic indexing for image retrieval systems. In *SIAM Conference on Applied Linear Algebra (LA03)*, 2003.

[21] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 1988.

[22] D. Silver and X. Wang. Tracking features in unstructured datasets. In *Proceedings of IEEE Visualization'98*, 1998.

[23] 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.

[24] 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.

[25] J. van Horn, J. Grethe, P. Kostelec, J. Woodward, J. Aslam, D. Rus, Rockmore, and M. Gazzaniga. The functional magnetic resonance imaging data center (fmridc): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philos Trans R Soc Lond B Biol Sci*, 356:1323–1339, 2001.

[26] I. Witten, A. Moffat, and T. Bell. *Managing gigabytes.* Van Nostrand Reinhold, 1994.

[27] 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.