# Effectiveness of the Finite Impulse Response Model in Content-Based fMRI Image Retrieval*

Bing Bai[1], Paul Kantor[2], and Ali Shokoufandeh[3]

[1] Department of Computer Science, Rutgers University
bbai@cs.rutgers.edu
[2] Department of Library and Information Science, Rutgers University
kantor@scils.rutgers.edu
[3] Department of Computer Science, Drexel University
as79@drexel.edu

**Abstract.** The thresholded t-map produced by the General Linear Model (GLM) gives an effective summary of activation patterns in functional brain images and is widely used for feature selection in fMRI related classification tasks. As part of a project to build content-based retrieval systems for fMRI images, we have investigated ways to make GLM more adaptive and more robust in dealing with fMRI data from widely differing experiments. In this paper we report on exploration of the Finite Impulse Response model, combined with multiple linear regression, to identify the "locally best Hemodynamic Response Function (HRF) for each voxel" and to simultaneously estimate activation levels corresponding to several stimulus conditions. The goal is to develop a procedure for processing datasets of varying natures. Our experiments show that Finite Impulse Response (FIR) models with a smoothing factor produce better retrieval performance than does the canonical double gamma HRF in terms of retrieval accuracy.

## 1 Introduction

As a method for watching "how the brain works", fMRI has become a powerful research tool in many aspects of neuroscience studies in the past decade [1]. More recently, classification of fMRI images, based on similarity between activation patterns, shows promising transition to clinical diagnosis [2,3,4]. These methods usually select features (that is to say, voxels or areas in the brain and their activation levels) and train models to best distinguish uncommon cases from so-called "typical" ones.

We investigate *content-based* indexing of fMRI images. For any "query" fMRI image that is presented, we ask whether we can retrieve images that represent the same or similar cognitive processes ("success"). The potential applications include, but are not limited to, the following: 1) helping doctors to diagnose brain disorders, by looking at the clinical history of persons with similar fMRI

* This work is supported by National Science Foundation Grant ITR-0205178. We thank Sven Dickinson, Deborah Silver and Nicu Cornea for significant discussions.

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part II, LNCS 4792, pp. 742–750, 2007.

patterns; 2) helping researchers to find similar studies and related research work; 3) helping researchers to discover hidden similarities among superficially different cognitive activities.

Our experimental studies are performed in the framework of information retrieval (IR) [5]. This framework is best known for applications such as search engines, which usually have a huge database of documents and images. In an IR framework, as in classification tasks, a dataset is represented in terms of a set of features. However, the IR framework is usually built to retrieve similar datasets from a very large database, in which it is generally difficult to assign class labels to each dataset. In contrast to seeking "class boundaries" optimized for specific classes, IR techniques use a more general "distance" measure. The IR framework is more extensible, and thus is preferable for an anticipated large database of fMRI datasets from miscellaneous sources.

In recent few years, a number of papers have been published on content-based fMRI retrieval [6,7,8]. These papers present matching methods based on features selected using General Linear Model (GLM) [9] t-maps.

In this paper, instead of testing matching methods with given features, we explore the possible ways to provide better features. Particularly, we find that inaccuracies in the assumed Hemodynamic Response Function (HRF), or in the associated stimulus time series may increase error in feature selection, and undermine the precision of subsequent processing. For example, in an experimental study of morality and decision-making [10], the subject presses a button when he/she thinks there is a moral issue to be resolved. In the reported analysis of this data, the beginning of the process of "moral reasoning" is set to be 8 seconds before the button is pressed, and the duration of this "stimulus" is set to 16 seconds. This approach works well with the specific method used in [10], but we find it can not be used in conjunction with general linear model in other typical settings [7]. In dealing with large heterogeneous data collections, we would not be able to generate either specialized HRF or stimulus configurations. Instead, we seek an adaptive HRF model, robust in handling cognitive processes with poor time definition, and efficient enough to allow large scale data processing.

The contributions of this paper can be summarized as follows. Firstly, we investigate the smoothing given by the Maximum A Posteriori (MAP) FIR model [11] as an adaptive HRF model for feature selection. This model exhibits better results in our experimental evaluations on real data than does the canonical HRF model. Secondly, we have extended this MAP FIR model to support multiple stimulus conditions, and propose a bilinear regression approach. This work has potential to be developed in a number of ways and the preliminary results show that it merits further study.

## 2 Method

### 2.1 GLM Based Feature Selection Schemes

In the GLM, observations (the time dependence of the signal at each voxel) $\mathbf{y}$ are to be explained by an intermediating variable $\mathbf{X}$ as $\mathbf{y} = \mathbf{X}\mathbf{b}$. $\mathbf{X}$ denotes the design matrix, and every one of its columns is an "Explanatory Variable"(EV) generated by convolving a "condition Stimulus" time series with an HRF. A popular choice for the HRF is the so-called "Canonical HRF" [1], which should be represented as the difference of two gamma functions: $H(t) = f(t; 6, 1) - \frac{1}{6}f(t; 16, 1)$, where $f(t; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-t/\beta}$ for $t > 0$. For each voxel, a t-value can be calculated, which indicates the significance of the voxel's activation by the corresponding condition. A 3D image of these t-values for all brain voxels will be referred to as a "t-map".

The first step of our feature selection scheme is based on the construction of t-maps and will be followed by selection of subset of the voxels with highest t-values. One straightforward idea is to set a threshold for t-values themselves, and take all voxels above this threshold as the features. Despite its superficial attractiveness, large variation of t-values of fMRI images for a large database of diverse experiments will make this mechanism unusable. In our database, for example, the maximum t-value is only about 3 for some experiments, while others can have t-values larger than 10, making it hard to set a reasonable threshold for all experiments. In our experiments, therefore, we uniformly select 1% of the voxels with the most significant t-values. We choose this "magic number" of 1 percent for two reasons: 1) indexing large databases calls for small feature sets, and 2) our main objective is to construct a robust HRF for information retrieval purposes, not to set some optimal threshold. In fact, we tried several different thresholds which resulted in similar relationships among different HRF models.

### 2.2 Finite Impulse Response (FIR) Model

Despite its simplicity, the canonical HRF model obviously fails to allow variations across multiple subjects or multiple brain regions. Temporal derivatives of EVs are sometimes included in design matrix to address very minor time shift [12], but the timing errors in real data may be much larger. As an alternative, more flexible models such as the Finite Impulse Response (FIR) have been proposed [11]. In these models, the activation of a certain voxel at time $t$ is the weighted sum of the stimulus values ($s_i, i \in [t - n + 1, t]$) at the preceding $n$ time points, i.e., $\hat{y}_t(\mathbf{w}) = \sum_{i=1}^{n} w_i s_{t-(i-1)} + w_0$.

The optimal estimate of $\mathbf{w} = [w_0, w_1, w_2, ...w_n]^T$ is taken to minimize the total squared error between the observations and the model. To avoid overfitting problems, Goutte et al. [11] adopted a maximum *a posteriori* (MAP) parameter estimation similar to ridge regression, $\mathbf{w}_{MAP} = (\mathbf{S}^T\mathbf{S} + \sigma^2\Sigma^{-1})^{-1}\mathbf{S}^T\mathbf{y}$, where $\Sigma_{ij} = v \exp(-\frac{h}{2}(i-j)^2)$, $h$ is a smoothing factor, $v$ is the strength, and $\sigma^2$ is the variance of noise [11]. Such a smoothing induces a correlation among parameters and prevents sudden changes (spikes) in the local form of the HRF. We shall refer to this model as "MAP FIR" in the rest of this paper.

### 2.3 FIR Model for Multiple Conditions at the Same Time

The aforementioned FIR model can only deal with a single stimulus condition. However, it is quite common that several conditions occur in a single fMRI run.

Although we could deal with this by using each condition separately in single regression, there is a potential problem with that approach. Suppose several conditions have similar effects on one voxel. If we consider only one condition, then the residual sum of squares $RSS$ will be greater in comparison to considering all conditions simultaneously, and this results in a smaller t-value. In other words, voxels whose time series are in fact just noise have a better chance to be selected. An example is shown in Figure 1.

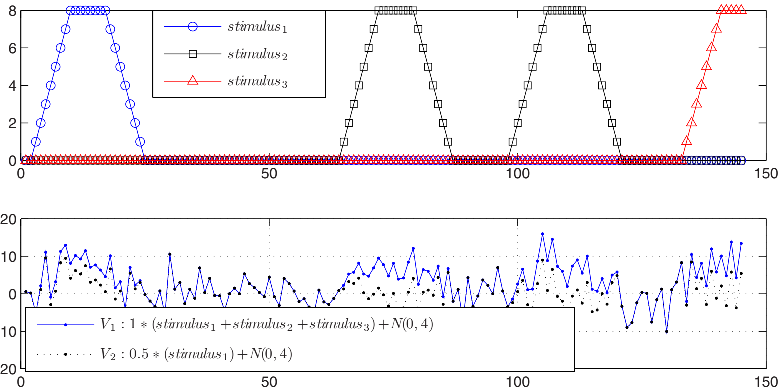

**Fig. 1.** Upper: 3 conditions, marked as $stimulus_1$,$stimulus_2$ and $stimulus_3$. Lower: voxel $V_1$ responds to all 3 stimuli with strength 1, while voxel $V_2$ to only the first stimulus, with strength 0.5. Both of them are subject to the same noise time series of N(0,4).

In Figure 1, we apply the GLM with multiple conditions or single condition to these two voxels, and inspect their t-values for condition $stimulus_1$. As shown in Table 1, the two methods select different voxels. For multiple regression, the t-value of $V_1$ is greater than $V_2$; for single regression it is the other direction. This is because, in single regression, the two other stimuli are considered as noise, lowering the confidence level associated with $stimulus_1$.

**Table 1.** t-values for $stimulus_1$ on voxels $V_1$ and $V_2$, with multiple regression and single regression respectively. (Generated with SPSS 11.5).

| | $V_1$ | $V_2$ |
|---|---|---|
| Multiple regression | **6.983** | 3.636 |
| Single regression | 3.570 | **3.698** |

Based on these observations, we propose to combine FIR model with multiple regression, and explore its effect in retrieval performance. This, in turn, allows us to simultaneously compute estimates for the HRF and for the activation levels. Specifically, we will assume that the *shape* of HRF is the same for different stimuli, at a given voxel, because the HRF describes a physiological feature of certain brain region, and that should not depend on how much the region is engaged in a process, nor on why it is engaged.

Specifically, suppose we have $c$ conditions, whose stimulus time series are: $s_j^i, i \in [1, c], j \in [1, N]$. Then an estimate for the activation at time $t$ can be written in the following parametric form:

$$\begin{aligned}\hat{y}_t &= \sum_{j=1}^{c} a_j \sum_{i=1}^{n} w_i s_{t-(i-1)}^j = \mathbf{a}^T \mathbf{S_t} \mathbf{w} \\ &= (a_1, a_2, ...a_c) \begin{pmatrix} s_t^1 \; s_{t-1}^1 \; ... \; s_{t-(n-1)}^1 \\ s_t^2 \; s_{t-1}^2 \; ... \; s_{t-(n-1)}^2 \\ ... \quad ... \quad ... \quad\quad ... \\ s_t^c \; s_{t-1}^c \; ... \; s_{t-(n-1)}^c \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ ... \\ w_n \end{pmatrix} \end{aligned} \tag{1}$$

For clarity, we omitted the constant terms from Eq. 1. The optimal value for the entries of weight vector **a** and the HRF **w** can be found using an “alternating” regression. That is, we fix **a** and **w** alternately, and calculate the other using linear regression until the process converges, as shown in Algorithm 1.

**Algorithm 1.** BilinearRegression($S$,$Y$,$MAP$) Iteratively find HRF and weights of regressors using alternating regression

```
$\mathbf{a} \leftarrow \mathbf{0}$; $\mathbf{a_{old}} \leftarrow -\infty$; $\mathbf{w} \leftarrow \mathbf{1}$
$iterations \leftarrow 0$
Build $SS$ from $S$
while $\|\mathbf{a} - \mathbf{a_{old}}\|_2 >$ NormThres and $iterations <$ IterThres do
    /* Estimate $a$ using $w$*/
    $U \leftarrow (\mathbf{S_1 w}, \mathbf{S_2 w}, \ldots, \mathbf{S_N w})^T$
    $\mathbf{a_{old}} \leftarrow \mathbf{a}$
    $\mathbf{a} \leftarrow (U^T U)^- U^T Y$
    /* Estimate $w$ using $a$*/
    $V \leftarrow (\mathbf{S_1}^T \mathbf{a}, \mathbf{S_2}^T \mathbf{a}, \ldots, \mathbf{S_N}^T \mathbf{a})^T$
    if $MAP$ then
        $\mathbf{w} \leftarrow (V^T V + var\Sigma^{-1})^- V^T Y$
    else
        $\mathbf{w} \leftarrow (V^T V)^- V^T Y$
    end if
    $iterations \leftarrow iterations + 1$
end while
return $\mathbf{w}$
```

This algorithm is guaranteed to converge, because linear regressions always reduces the least square error $\sum_{t=1}^{N}(y_t - \hat{y}_t)^2$, which is non-negative. With respect to landscape of local and global minima, the convergence behavior is not completely clear at this moment. However, in our validating experiments, we found that longer voxels time series and fewer conditions yield fewer local minima.

## 3 Results

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 condition. It is possible, of course, that data with different labels may contain similar brain process. In this case, the hidden similarity across conditions will increase the rank of items considered irrelevant, and lower the retrieval performance metric. Thus, the metric that we calculate should be a lower bound for the retrieval based on *real* similarity (including similarities not yet known to cognitive scientists). See [7] for more details about this framework.

Since the number of examples for each condition may be quite different, we choose a metric insensitive to data size. We use the "Area Under the ROC Curve" (AUC) to evaluate each retrieval method. If the AUC is 0.5, then the retrieval method is no better than random selection. An AUC of 1 is a perfect retrieval. We use each of the datasets as a query against the rest (excluding the same subject), calculate AUC for each ranked list, and report the average AUC of all queries as the performance indicator.

The similarity measure we use between two thresholded t-maps is the Jaccard distance. Specifically, the similarity between two sets of selected voxels is simply the size of their overlap divided by the size of their union, $similarity(A, B) = \|A \cap B\| / \|A \cup B\|$. The hyper parameters in MAP FIR model are $h = .3$, $v = .1$, and $\sigma^2 = 1$.

We have gathered 430 real fMRI datasets from different institutions. Table 2 shows details of this testing database. These data are preprocessed (motion correction, spatial smoothing, high pass filtering, and registration to standard brain space) with the software package FSL [13]. To eliminate artifacts introduced by the fact that different brain regions are scanned in different experiments, we specifically consider only those parts of the brain that were scanned in *all* images. This is similar to the approach used in Mitchell et al. [4].

Our study explores the combination of single or multiple regression, with the canonical or finite impulse models for the HRF. Table 3 shows the average AUC for four different combinations of these two aspects. "CAN", "MAP", "SIN",

**Table 2.** Experiments

| **Experiment** | **Conditions** | **TR(s)** | **Size** |
|---|---|---|---|
| Oddball: Recognition of an out of place image or sound | auditory, visual | 2.0 | 8 |
| Event perception: Watching either a cartoon movie of geometric shapes or real film of a human being [14] | studyActive, houseActive | 1.5 | 53 |
| Morality: Making decisions about problem situations having or lacking combinations of moral and emotional content [10] | M+E+, M+E-, M-e- | 2.0 | 150 |
| Recall: Study and recall or recognition of faces, objects and locations [15] | {S,T,R}{Face, Obj, Loc} | 1.8 | 189 |
| Romantic: People in love seeing pictures of their significant others, or of non-significant others [16] | neutralFace, positiveFace | 5.0 | 30 |

and "Mul" denoting "Canonical HRF", "MAP HRF", "Single regression", and "Multiple regression", respectively. "AAUC (raw)" is the Average AUC for all 430 queries. Since this metric tends to be dominated by conditions with many samples, we also calculate the mean AUCs for each condition, and refer to the mean value of those as the "(Macro-)adjusted AUC".

**Table 3.** Average AUC for 430 datasets (Mean/Standard Error of the mean)

| | CAN_MUL | MAP_MUL | CAN_SIN | MAP_SIN |
|---|---|---|---|---|
| AAUC (raw) | .662/.007 | **.719/.006** | .677/.007 | .715/.007 |
| AAUC (adjusted) | .658/.006 | .711/.005 | .665/.007 | **.715/.006** |

We test two hypotheses using these results. **H1**: "the FIR model performs better than canonical HRF in retrieval". The hypothesis is clearly accepted since the differences are very significant for both single-variate and multi-variate approaches. **H2**: "for series of brain scans with multiple conditions, one multiple regression with all conditions performs better than multiple simple regressions". The conclusion for this hypothesis is not clear yet. Figure 2, which shows the AAUC for separate conditions (see Table 2), provides further detail on this. Each method is better for *some* of the conditions. We return to this point in the discussion.

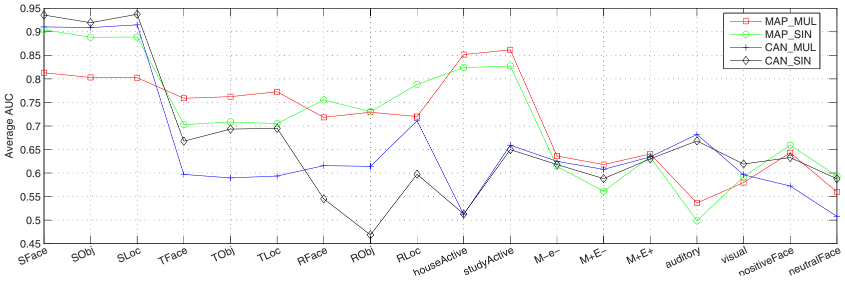

**Fig. 2.** Average of area under ROC (RAW) curve for 4 methods

## 4 Conclusions and Discussions

The results of this study are: confirmation of one hypothesis (H1) , and some tantalizing clues regarding the other. Specifically, the FIR model, with MAP smoothing, which seems to be a more realistic way to describe the variations, across the brain, in the anatomy supplying blood, *does* also yield significantly better performance in the retrieval setting. This suggest that it may be worth the added effort to use smoothed FIR analysis when preparing data for retrieval across different experiments, and different laboratories.

On the other hand, the anticipated superiority of using multiple independent regressors to select the voxels characteristic of several cognitive conditions occurring in the same run, is not confirmed. This lead us to a more detailed examination of *why* it was expected to be better, and a new hypothesis.

Our argument in favor of using multivariate regression relied on the assumption that an individual voxel may be activated by several conditions, all occurring in the same experimental run. Using all but the condition of interest as a contrast has the effect of making the estimates of correlation with the signal less accurate. This makes the t-value smaller, and makes the voxel less likely to be selected as a feature. On the other hand, conditions that activate same voxels are harder to tell distinguish the same run. One or the other of these two contradictory factors may dominate in different experiments. As shown in Figure 2, for some types of experiments the multivariate regression (e.g., M+E+, M+E- and M-e-) *is* more effective, while for some of them (e.g. SFace, SLoc and SObj) it is not. This relationship will be further investigated in future work.

Another interesting topic is the distribution of estimated FIR weights. It can not be included here due to page limit. Please see [17] for a brief report.

## References

1. Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W.: Human Brain Function, 2nd edn.
2. Ford, J., Farid, H., Makedon, F., Flashman, L., McAllister, T., Megalooikonomou, V., Saykin, A.: Patient classification of fMRI activation maps. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2878, Springer, Heidelberg (2003)
3. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fMRI data. NeuroImage 26, 317–329 (2005)
4. Mitchell, T., Hutchinson, R., Pereira, R.N., Wang, F.: Learning to decode cognitive states from brain images. Machine Learning 57, 145–175 (2004)
5. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513 (1988)
6. Bai, B., Kantor, P., Cornea, N., Silver, D.: IR principles for content-based indexing and retrieval of functional brain images. In: Proceedings of the CIKM 2006 (2006)
7. Bai, B., Kantor, P., Cornea, N., Silver, D.: Toward content-based indexing and retrieval of functional brain images. In: Proceedings of the RIAO 2007 (2007)
8. Zhang, J., Megalooikonomou, V.: An effective and efficient technique for searching for similar brain activation patterns. In: Proceedings of the ISBI 2007 (2007)
9. Friston, K., Jezzard, P., Turner, R.: Analysis of functional MRI time-series. Human Brain Mapping 1, 153–171 (1994)
10. Greene, J., Sommerville, R., Nystrom, L., Darley, J., Cohen, J.: An fMRI investigation of emotional engagement in moral judgment. Science 293 (2001)
11. Goutte, C., Nielsen, F.Å., Hansen, L.K.: Modelling the haemodynamic response in fMRI with smooth FIR filters. IEEE Trans. Med. Imaging 19(12), 1188–1201 (2000)
12. Smith, S.: Overview of fMRI analysis. The British Journal of Radiology (77), S167–S175
13. Smith, S., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, M., Leibovici, D., Ripley, B., Woolrich, M., Zhang, Y.: FSL: New tools for functional and structural brain image analysis. In: Seventh Int. Conf. on Functional Mapping of the Human Brain. NeuroImage vol. 13, p. S249 (2001)

14. Zaimi, A., Hanson, C., Hanson, S.: Event perception of schema-rich and schema-poor video sequences during fMRI scanning: Top down versus bottom up processing. In: Proceedings of the Annual Meeting of the Cognitive Neuroscience Society (2004)
15. Polyn, S., Cohen, J., Norman, K.: Detecting distributed patterns in an fMRI study of free recall. In: Society for Neuroscience conference (2004)
16. Aron, A., Fisher, H., Mashek, D., Strong, G., Li, H., Brown, L.: Reward, motivation, and emotion systems associated with early-stage intense romantic love. J. Neurophysiol. 94, 327–337 (2005)
17. Bai, B., Kantor, P.: A shape-based finite impulse response model for functional brain images. In: Proceedings of the ISBI 2007 (2007)