Modeling the Brain Connectivity for Pattern Analysis

Itir Onal*, Emre Aksan*, Burak Velioglu*, Orhan Firat*, Mete Ozay†, Ilke Oztekin‡, Fatos T.Yarman Vural*
*Department Computer Engineering, Middle East Technical University, Ankara, Turkey
Emails: {itir,emre.aksan,velioglu,orhan.firat,vural}@ceng.metu.edu.tr
†School of Computer Science, University of Birmingham, Birmingham, UK
Email: m.ozay@cs.bham.ac.uk
‡Department of Psychology, Koc University, Istanbul, Turkey
Email: ioztekin@ku.edu.tr

***Abstract*—An information theoretic approach is proposed to estimate the degree of connectivity for each voxel with its neighboring voxels. The neighborhood system is defined by spatial and functional connectivity metrics. Then, a local mesh of variable size is formed around each voxel using spatial or functional neighborhood. The mesh arc weights, called *Mesh Arc Descriptors (MAD)*, are estimated by a linear regression model fitted to the voxel intensity values of the functional Magnetic Resonance Images (fMRI). Finally, the error term of the linear regression equation is used to estimate the mesh size for a voxel by optimizing Akaike's information Criterion, Bayesian Information Criterion and Rissanen's Minimum Description Length. fMRI measurements are obtained during a memory encoding and retrieval experiment performed on a subject who is exposed to the stimuli from 10 semantic categories. For each sample, a k-NN classifier is trained using the *Mesh Arc Descriptors (MAD)* having the variable mesh sizes. The classification performances reflect that the suggested variable-size *Mesh Arc Descriptors* represents the mental states better than the classical multi-voxel pattern representation. Moreover, we observe that the degree of connectivities in the brain greatly varies for each voxel.**

I. Introduction

Recent studies on modeling and analysis of fMRI data employ the full spatial pattern of brain activity and use pattern classification algorithms to decode the subtle information represented in a cognitive state [1]. Identification of patterns that are predictive of cognitive states and using them in classification are called Multi-voxel pattern analysis (MVPA). A pioneering study on MVPA methods to decode the cognitive states from the fMRI data was conducted by Haxby et al. [2]. Since then, many studies [3], [4], [5], [6], [1], [7] employed MVPA for various cognitive state classification problems. In classical MVPA approaches, generally, cognitive states are represented by concatenating the voxel intensity values under a feature vector and training a well-known classifier such as k-Nearest Neighbor (k-NN), Support Vector Machine (SVM) or Naive Bayes.

Recently there has been interest in decoding brain states using fMRI in pattern recognition community. Plumpton et al. [8], [9] proposed two pattern recognition approaches for online classification of fMRI data. While the former uses linear and ensemble classifiers to decode cognitive states, the latter performs a semi-supervised ensemble update strategy. Moreover, Gramfort et al. [10] used supervised learning methods to decode the visual percept formed by four letters during a word reading task. In their study Haufeld et al. [11] employed various kinds of supervised self-organizing maps (SSOM) to decode and visualize fMRI voxel patterns assigned to multiple categories.

In [12], Ozay et al. propose a local mesh model in which relationships among *spatially close* voxels are used as features. These features, called *Mesh Arc Descriptors*, are shown to discriminate the cognitive states better than voxel intensities. In a further study, Firat et al. [13] model the relationships among *functionally close* neighbors and use the arc weights of functionally connected voxels to train a classifier. These studies show a significant improvement on the performance of the algorithms developed for cognitive state classification. In both studies, the number of neighboring voxels is fixed to form the mesh assuming that each voxel is connected to the same number of neighboring voxels to represent a cognitive process. In [14], [15], Onal et al. adopt a set of information theoretic criteria to estimate the optimal mesh size for each participant and sample. Although these approaches reflect the distributivity of information in the brain for an individual participant and sample, they form the local mesh of the same size around each voxel belonging to a sample. However, as Baldassano et al. [16] states, different sub-regions of brain have different degree of sub-connectivities among the voxels. Furthermore, according to Zalesky et al. [17], both the topology and the strength of connectivities of brain networks differ in different regions.

In this study, we propose an information theoretic approach to model the spatial and functional brain connectivity among the voxels for a cognitive process. The degree of connectivity of a voxel is represented by the ”optimal” size of a local mesh, formed around each voxel. The size of the mesh is estimated by optimizing three well-known information theoretic criteria namely Akaike Information Criterion (AIC) [18], Bayesian Information Criterion (BIC) [19] and Minimum Description Length (MDL) [20]. Unlike previous approaches, this study aims to form meshes of variable sizes to detect how the information distribution varies in different parts of the brain. The local meshes are formed in two neighborhood systems, where nearest neighbors of a seed voxel are either the ones that are spatially closest to the seed voxel, or the ones whose Pearson correlations are highest among others.

The *Mesh Arc Descriptors* extracted from meshes of variable sizes are used to train a k-NN classifiers to classify cognitive states. Our classification performances indicate that the proposed spatial and functional brain connectivity models represent the cognitive states with a higher accuracy than classical MVPA methods.

II. fMRI Experiment and Dataset

Our dataset consists of words belonging to 10 categories, namely, fruits, vegetables, furniture, animals, herbs, clothes, body parts, chemical elements, colors and tools. In the experiments, a participant is shown a list of words belonging to a specified category in the encoding phase. fMRI measurements recorded in this phase are used to train the classifier. Then, the participant solves mathematical problems in the delay period. Finally, in the retrieval phase, the participant is shown a word belonging to the same specified category and is expected to recognize whether the word is included in the list or not [21], [22]. The fMRI measurements recorded during the retrieval phase are used to test the classifier. Therefore, training data are collected during the encoding phase whereas test data are collected during the retrieval phase. The experiments are repeated in 8 runs. The region of interest (ROI) is the lateral temporal cortex with 8142 voxels that reside in this ROI.

In this study, fMRI intensity of a voxel at coordinates $\bar{s}_j$, measured at time instant t_i, is represented as $v(t_i, \bar{s}_j)$ where $\bar{s}_j$ is a three dimensional coordinate vector $\bar{s}_j = (x_j, y_j, z_j)$. Note that, $i = 1, 2, ..., N$ and $j = 1, 2, ..., M$, where N is the number of samples and M is the number of voxels. The fMRI intensities measured at all voxel coordinates $\bar{s}_j$ for a single time instant t_i are used to form a $1 \times M$ vector, which is called a *sample* vector. By concatenating each *sample* vector, we form our dataset $D\{v(t_i, \bar{s}_j)\}$ of size $N \times M$. During our experiments, each *sample* measured at time instant t_i is assigned a class label c_i.

III. Mesh Arc Descriptors (MAD)

In their study Ozay et al. [12] proposed a mesh model to classify cognitive states, in which a local mesh is formed around a seed voxel $v(t_i, \bar{s}_j)$ with its *p-nearest neighbors* $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$.

In this study, we form two types of local meshes, depending on the neighborhood system, η_p. The first neighborhood system is based on the spatial distance between the voxels, whereas the second system measures the functional similarity between the time series of the voxels. In order to form the local meshes with respect to spatially *p-nearest neighborhood*, p number of voxels having the smallest Euclidean distance between the coordinates of the seed voxel and its neighbors are selected in the mesh of the voxels, $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ [12]. On the other hand Firat et al. [13] defined the *p-nearest neighborhood* functionally, where *p-nearest neighbors* are selected based on the functional connectivity between the seed voxel and the surrounding voxels. In this approach *p-nearest neighbors* $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ of a seed voxel are the ones where the Pearson correlations with the seed voxel are the highest p voxels.

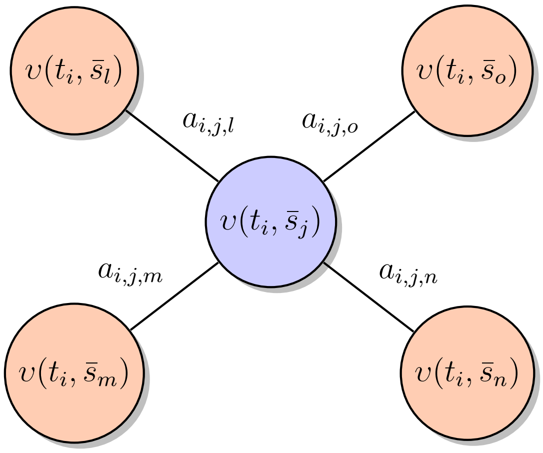

Fig. 1: A local mesh representing the relationships among the seed voxel $v(t_i, \bar{s}_j)$ and its *p-nearest neighbors* $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ with the arc weights $a_{i,j,k}$

Figure 1 shows a local mesh formed around a seed voxel. While the voxel intensity values are represented as the vertices of the local mesh, the relationships between the seed voxel $v(t_i, \bar{s}_j)$ and its *p-nearest neighbors* $\{v(t_i, \bar{s_k})\}_{k=1}^{p}$ are represented with the arc weights $a_{i,j,k}$ between the corresponding vertices. The arc weights $a_{i,j,k}$ which represent the of relations between a voxel and its neighbors are called *Mesh Arc Descriptors (MAD)*. During theexperimental analysis, we observe that the relationship between a seed voxel and its neighbors are very close to linear. Therefore, we model each voxel as a linear combination of its neighbors and estimate the MAD features using the following equation:

$$v(t_i, \bar{s}_j) = \sum_{\bar{s}_k \in \eta_p} a_{i,j,k} v(t_i, \bar{s}_k) + \varepsilon_{i,j}(p) \qquad (1)$$

where the seed voxel is $v(t_i, \bar{s}_j)$, the *p-nearest neighbors* are $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$. $\varepsilon_{i,j}(p)$ is the error obtained during the estimation of the MAD features $a_{i,j,k}$ of local mesh formed at the time instant t_i and for the seed voxel having coordinates $\bar{s}_j$.

There are many ways of estimating the MAD features. One popular method is to minimize the regularized least squares error of $\varepsilon_{i,j}(p)$, using Ridge regression [ref itir]. In other words we minimize

$$L = \varepsilon_{i,j}(p) + \lambda \sum_{\bar{s}_k \in \eta_p} a_{i,j,k}. \qquad (2)$$

Another method is to order the neighbors of each voxel from the most similar to least similar ones and represent each mesh as a one dimensional linear predictive model. In this study, we employ both Ridge regression and linear predictive model and observe that the estimated parameters are very similar to each other. For this reason , we provide the results only obtained from the linear predictive coding [23] via Levinson-Durbin recursion, $lpc(d, p)$ **[?]** , which minimizes the variance of error $\varepsilon_{i,j}(p)$ of (1). Note that in $lpc(d, p)$, d is a 1-dimensional series, starting at the *seed voxel* and ending at the *pth nearest neighboring voxel*, which is sorted by the distance between the seed voxel and its neighbors. Therefore, $d = \{v(t_i, \bar{s}_j) \cup \{v(t_i, \bar{s}_k)\}_{k=1}^{p}\}$ and p is the mesh size.

Mesh Arc Descriptors are, then, used to form a mesh arc vector $\bar{a}_{i,j} = [a_{i,j,1}, a_{i,j,2}, ...a_{i,j,p}]$ of size $1 \times p$ around voxel $v(t_i, \bar{s}_j)$. By combining the mesh arc vectors for all voxels at time instant t_i, we form mesh arc vector for *sample* at t_i, $A_i = [\bar{a}_{i,1}, \bar{a}_{i,2}, ...\bar{a}_{i,M}]$ having size $1 \times p.M$. Finally, mesh arc vectors for all *samples* belonging to a participant are combined to form the feature matrix $F = [A_1^T, A_2^T, ...A_M^T]^T$ of size $Nxp.M$.

IV. Information Criteria to Estimate the Degree of Voxel Connectivities

The size p of a local mesh defines the number of neighboring voxels, which are connected to the seed voxel. In other words, p defines the degree of connectivity at a specific location of the brain. If the mesh size p is large, the seed voxel makes dense connections with its neighbors whereas a small p value indicates a sparse connection. As the mesh size p increases, the error term $\varepsilon_{i,j}(p)$ of (1) decreases and the model fits the data better. On the other hand, the mesh size p is equal to the number of mesh arc descriptors extracted from a mesh. Since the mesh arc descriptors are used as features in our system, the complexity of our system increases with an increase in p. The trade-off between the degree of fit and complexity can be optimized by some information theoretic criteria. In order to find the optimal mesh size, we adopt three information theoretic criteria namely Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Minimum Description Length (MDL). In this study, we applied the same criteria to estimate the optimal sizes of both spatially and functionally formed local meshes.

In order to optimize the above mentioned information criteria, first, the squared error of (1) is calculated for each voxel and for each mesh size, as follows,

$$\varepsilon_{i,j}\left(p\right)^{2} = \left(v\left(t_i, \bar{s}_j\right) - \sum_{\bar{s}_k \in \eta_p} a_{i,j,k} v\left(t_i, \bar{s}_k\right) \right)^2 . \quad (3)$$

In this study the expected value of squared error is approximated for a seed voxel $v(t_i, \bar{s}_j)$ by averaging the errors obtained during the formation of a local mesh around the seed voxel $v(t_i, \bar{s}_j)$ over all time instants as

$$\hat{E}_j\left(p\right) = E_i\left(\varepsilon_{i,j}(p)^2\right) \cong \frac{1}{N}\sum_{i=1}^{N} \varepsilon_{i,j}(p)^2 , \quad (4)$$

where $E_i(\cdot)$ is the expectation operator taking the average over all time instants t_i for a single voxel.

Our goal is to estimate the optimal mesh size for each voxel using the error variance of equation 3. Then, this variable size of local meshes reflect how the information distribution varies around each voxel.

A. Akaike Information Criterion (AIC)

If the information distribution in the brain were known, we would compute the Kullback - Leibler (KL) divergence between the true model and our local mesh model of size p and select the order which minimizes KL. Therefore, we would measure the information loss of using our model. Since the information distribution in the brain is unknown, we can approximate it by using Akaike Information Criterion (AIC). The optimal mesh size estimated using AIC is the one that leads to minimum information loss during the approximation of distribution in the brain.

In order to estimate the optimal mesh size for the voxel at coordinates $\bar{s}_j$ with AIC [18], the following criterion is minimized with respect to the mesh size p:

$$AIC_j(p) = \ln(\hat{E}_j) + \frac{2p}{M} \quad (5)$$

where M is the number of voxels. After calculating the $AIC_j(p)$ for various mesh sizes, the optimal mesh size for voxel at coordinates $\bar{s}_j$ is selected as the one minimizing $AIC_j(p)$,

$$\hat{p}_j^{AIC} = \underset{p \in [p_{min}, p_{max}]}{\arg\min} \left(AIC_j(p)\right) \quad (6)$$

where $\hat{p}_j^{AIC}$ is the optimal mesh size for voxel at coordinates $\bar{s}_j$ and $[p_{min}, p_{max}]$ is an interval determined empirically, in which the optimal mesh size lies.

B. Bayesian Information Criterion (BIC)

While AIC assumes that true model is unknown and can only be approximated, Bayesian Information Criterion (BIC) assumes that the true model is among candidate models (local mesh models of various sizes) and tries to detect it. BIC aims to point out the model having optimal mesh size p as the one maximizing the probability of generating the data. By assuming that the error terms have a normal distribution, we adopt the BIC $(BIC_j(p))$ to estimate the optimal mesh size for each voxel at coordinates $\bar{s}_j$ as follows;

$$BIC_j(p) = \ln\left(\hat{E}_j(p)\right) + \frac{p\ln(M)}{M} , \quad (7)$$

where M is the number of voxels. The mesh size p minimizing $BIC_j(p)$ among various others is selected as the optimal mesh size estimated using BIC as;

$$\hat{p}_j^{BIC} = \underset{p \in [p_{min}, p_{max}]}{\arg\min} \left(BIC_j(p)\right) , \quad (8)$$

where $\hat{p}_j^{BIC}$ refers to the optimal mesh size for voxel at $\bar{s_j}$, where the experiment is conducted using BIC.

C. Minimum Description Length (MDL)

Minimum Description Length (MDL) assumes that the best model is the represented with the smallest description length without high information loss. Therefore, MDL estimates the optimal mesh size as the one that leads to compact representation without a significant loss. As explained before, MDL optimizes a trade-off between the degree of fit (without high information loss) and the complexity (smallest representation). MDL is adopted to estimate the optimal mesh size for a voxel at coordinates $\bar{s_j}$ using the following equation:

$$MDL_j(p) = \hat{E}_j(p)\left(1 + \left(\frac{p+1}{M}\right)\ln\left(M\right)\right) , \quad (9)$$

where M is the number of voxels. After $MDL_{j,}(p)$ is estimated for the mesh sizes in the pre-defined interval $[p_{min}, p_{max}]$, the optimal mesh size for voxel at coordinates $\bar{s_j}$ is selected as the one minimizing $MDL_j(p)$, as follows;

$$\hat{p}_j^{MDL} = \underset{p \in [p_{min}, p_{max}]}{\arg\min} \left(MDL_j(p) \right) \; , \qquad (10)$$

where $\hat{p}_j^{MDL}$ refers to the optimal mesh size for voxel at $\bar{s_j}$ estimated using MDL.

V. Degree of Connectivity in the Brain

The above approach for estimating the optimal mesh size, formed around each voxel shows the degree of spatial and functional connectivities of a voxel and its neighborhood. The validity of the suggested approach requires a throughout study from the perspective of cognitive science. However, in this study, we suffice to use the suggested model for mental state classification and compare the classification performances of the suggested model to the classical MVPA methods. We expect that the performance of each model indicates the validity or representation power of it.

Fig. 2 represents the optimal mesh sizes estimated for each voxel using spatial (A) and functional (B) neighborhood. Moreover, corresponding histograms of optimal mesh sizes estimated using spatial (C) and functional (D) neighborhood are also represented. In Fig. 2, the nodes correspond to voxel coordinates in Lateral Temporal Cortex. The colors of nodes represent the degree of connectivity of voxel. In other words, color represents the optimal mesh size estimated for the voxel. Notice that, the darkest red color represents the largest optimal mesh size whereas the darkest blue represents the smallest optimal mesh size. Remaining colors in between are assigned to voxels based on their optimal mesh size in a scale [24]. As it can be seen, the voxels having similar number of connections with their neighbors tend to group together and the separation between the groups are more visible in spatial neighborhood. It is interesting to note that functional connectivity has a smoother histogram compared to the spatial connectivity. This observation may indicate that the voxels in lateral temporal cortex is more likely related to each other according to functional neighborhood, for the underlying cognitive processes.

VI. Image Processing

Before extracting the MAD features, the fMRI data is enhanced by a series of pre-processing operations. These operations are accomplished by some standard techniques using Statistical Parametric Mapping toolbox. The details of the pre-processing techniques can be found in [12]. Then, the data is normalized using standardized z-scores. GLM analysis is conducted to z-score maps, using a design matrix composed of 22 columns (1 column for bias, 1 column for scanner-drift, 20 columns for encoding and decoding of 10 semantic categories). Design matrix is convolved with a double-gamma hemodynamic response function, except first two columns. Using a Generalized Linear Model (GLM) the beta weights, betamap are estimated. GLM model is constructed and betamap values are estimated using libORF (www.ceng.metu.edu.tr/ e1697481/libORF.html). Therefore, in this study we employ the betamap parameters, instead of the raw voxel intensity values.

VII. Construction of MAD features

Optimal mesh size for each voxel is estimated using AIC, BIC and MDL from the training data. The MAD features having variable mesh sizes for various local meshes are concatenated to form the feature vector for a training *sample* of size $1 \times D_{AIC}$ where $D_{AIC} = \sum_{j=1}^{M} \hat{p}_j^{AIC}$, $1 \times D_{BIC}$ where $D_{BIC} = \sum_{j=1}^{M} \hat{p}_j^{BIC}$ and $1 \times D_{MDL}$ where $D_{MDL} = \sum_{j=1}^{M} \hat{p}_j^{MDL}$ using AIC, BIC and MDL respectively. Using the training feature vectors, and their corresponding class labels we train a k-nearest neighbor (kNN) classifier. When a new test sample is queried, it is converted into the same feature vector form as the training feature vectors and using the k-NN classifier, a class label is assigned to it. Note that we have two types of local meshes: The first type of local meshes are constructed with respect to the spatial neighborhood, whereas the second one is formed by using the functional neighborhood. Therefore the MAD features are obtained in two different neighborhood systems.

VIII. Classification using MAD features

The classification is applied to both functionally connected and spatially connected MAD features. Since our dataset includes 8 runs, we train and classify the model run-wise for functionally connected MAD features. By run-wise we mean that a connectivity matrix consisting of pairwise correlations among voxels is computed separately for each run and classification is also performed using functionally formed meshes from each run. Therefore, we train and test 8 classifiers for functionally formed meshes. For the spatially connected MAD features, we train the classifier over the entire runs, since the spatial connectivity is assumed to remain unchanged over the runs. We also test the performances run-wise for this case and compare the performances. In that case, classification is computed for each run separately for spatially formed meshes. Three classification schemes, used in our experiments can be summarized as follows:

S_1 **Spatial neighborhood, Whole data classification:** MAD features are extracted using spatially nearest neighbors. Single classifier is trained using MAD features of whole training data.

S_2 **Spatial neighborhood, Run-wise classification:** MAD features are extracted using spatially nearest neighbors. Different classifiers are trained for each run using MAD features of the training data of the related run.

S_3 **Functional neighborhood, Run-wise classification, Connectivity matrix from each run:** MAD features are extracted using functionally nearest neighbors. For each run, connectivity matrix representing the pairwise connectivities of voxels for all time instants in the related run is formed. Different classifiers are trained for each run using MAD features of the training data of the related run.

IX. Results

In our experiments, we empirically determine the interval of mesh sizes $[p_{min}, p_{max}]$ as [2, 100] in which the optimal

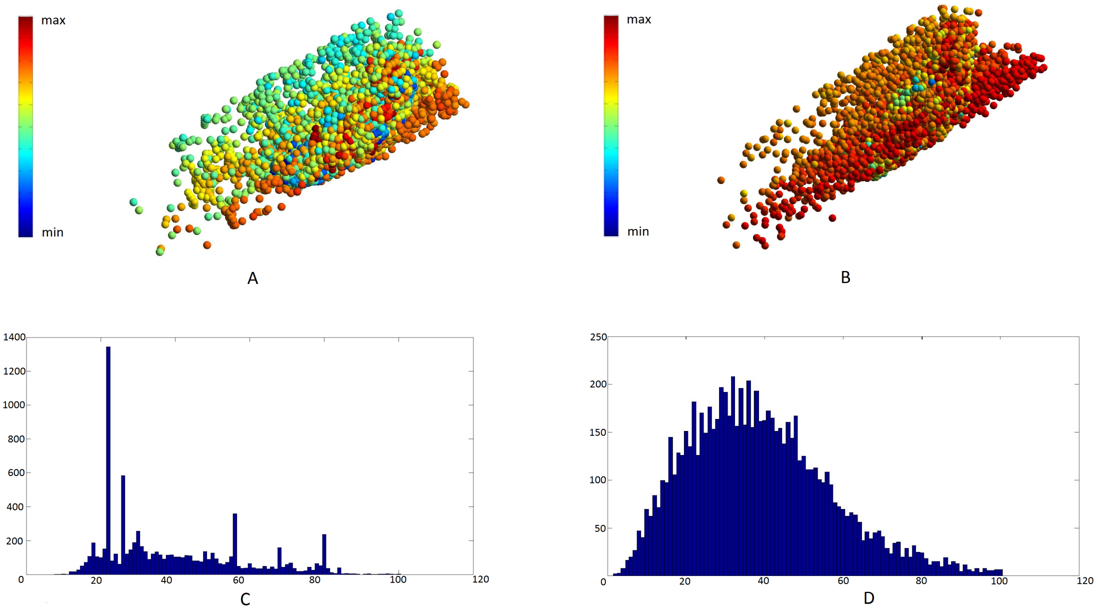

Fig. 2: Representation of optimal mesh sizes estimated for each voxel using spatial (A) and functional (B) neighborhood and histograms of optimal mesh sizes computed using spatial (C) and functional (D) neighborhood [24].

mesh sizes of all of the voxels are assured to lie. Since variable sizes of local meshes are formed around all voxels, we can calculate the mean (μ_{IC}) of optimal mesh sizes can be estimated as follows:

$$\mu_{IC} \cong \frac{1}{M} \sum_{j=1}^{M} \hat{p}_j^{IC} \ , \tag{11}$$

where IC is either AIC, BIC or MDL, M represents the number of voxels and $\hat{p}_j^{IC}$ is the optimal mesh size estimated for a voxel at coordinates $\bar{s_j}$. Similarly, standard deviation of optimal mesh sizes for all voxels can be calculated using:

$$\sigma_{IC} \cong \sqrt{\frac{1}{M} \sum_{j=1}^{M} (\hat{p}_j^{IC} - \mu_{IC})^2} \ . \tag{12}$$

Table I shows the mean and standard deviations of optimal mesh sizes calculated using AIC, BIC and MDL for the first 2 schemes.

TABLE I: Mean and standard deviations of optimal mesh sizes estimated using AIC, BIC and MDL in two different classification schemes S_1 and S_2

	AIC		BIC		MDL	
	μ_{AIC}	σ_{AIC}	μ_{BIC}	σ_{BIC}	μ_{MDL}	σ_{MDL}
$\boldsymbol{S_1, S_2}$	38.16	18.84	32.71	15.46	32.96	15.67

Since in the last scheme (S_3), the optimal mesh sizes are estimated for each run and different results are obtained, we can calculate the mean and standard deviations of optimal mesh sizes for each run. Table II reflects resulting mean and standard deviations run-wise.

TABLE II: Run-wise mean and standard deviations of optimal mesh sizes estimated using AIC, BIC and MDL in the last classification scheme S_3

	AIC		BIC		MDL	
	μ_{AIC}	σ_{AIC}	μ_{BIC}	σ_{BIC}	μ_{MDL}	σ_{MDL}
$\boldsymbol{run_1}$	48.59	26.77	43.87	24.64	44.25	24.93
$\boldsymbol{run_2}$	40.68	23.74	36.96	21.24	37.19	21.47
$\boldsymbol{run_3}$	40.77	24.76	37.34	22.36	37.58	22.60
$\boldsymbol{run_4}$	45.32	21.52	41.03	19.02	41.32	19.29
$\boldsymbol{run_5}$	44.43	23.59	40.30	21.07	40.61	21.33
$\boldsymbol{run_6}$	41.17	20.39	37.53	17.88	37.76	18.10
$\boldsymbol{run_7}$	46.46	22.70	41.96	20.09	42.32	20.39
$\boldsymbol{run_8}$	50.32	21.80	45.61	19.47	45.93	19.70

Notice that, Table I reflects standard deviations of mesh sizes estimated using spatial neighborhood whereas Table II includes the ones estimated using functional neighborhood for each run. From both tables it can be concluded that standard deviation of optimal mesh sizes is smaller in spatial neighborhood than in functional neighborhood.

Table III reflects classification performances of using three different schemes using variable size of p (S_1, S_2 and S_3) and fixed size of p (F_1, F_2 and F_3) in which optimal mesh size is estimated for participant as in [14]. Note that F_1 represents whole data classification where fixed size meshes are formed using spatial neighborhood, F_2 represents run-

TABLE III: Classification accuracies of using AIC, BIC and MDL in three different classification schemes

	AIC	BIC	MDL
$\boldsymbol{S_1}$	75%	74%	74%
$\boldsymbol{F_1}$	74%	71%	71%
$\boldsymbol{S_2}$	78%	75%	75%
$\boldsymbol{F_2}$	75%	71%	71%
$\boldsymbol{S_3}$	78%	78%	78%
$\boldsymbol{F_3}$	77%	77%	77%

wise classification where fixed size meshes are formed using spatial neighborhood and F_3 represents run-wise classification where fixed size meshes are formed using functional neighborhood. The classification performance of using classical MVPA method on this data, in other words training the classifier using the raw intensity values instead of MAD features performs 45% when k-NN is used as the classifier. Moreover, the average performance of training 8 classifiers for each run using MVPA method is 57%. As it can be seen from Table III, performance results of using our approaches are much higher than that of classical MVPA methods. Notice that, performing classification with the features extracted using BIC and MDL result in the same accuracies. Although using AIC performs slightly better than those two criteria in spatial neighborhood (S_1 and S_2), three criteria perform equally in functional neighborhood (S_3).

X. Conclusion

In this study, we propose an information theoretic approach to model the connectivity among the distributed patterns formed by the meshes of voxels in brain. First, we form a local mesh around each voxel, in two different neighborhood systems, by defining spatial and functional connectivity. Then, we estimate the mesh arc weights, called *Mesh Arc Descriptors* by a linear regression model, where the model order for each mesh is estimated by optimizing a set of information criteria. In this model, each voxel is connected to its p-spatially or functionally nearest voxels, where the p values vary for each voxel.

The main contribution of this study is to explore the functional and spatial connectivity among the voxels, during a cognitive process. In our study, the degree of connectivity is represented in terms of the optimal mesh size of a voxel which is estimated using various information theoretic criteria. Another contribution of this study is the employment of the MAD model to the classification of cognitive states. The validity of the suggested connectivity model is shown to a certain extent, by the improved classification performances of the cognitive states. Although there are slight differences in the performances, the proposed model has a higher classification accuracy than that of the classical MVPA methods. It can be concluded that using MAD with variable sizes has a better representational power to discriminate the cognitive states.

Anatomical and cognitive correspondences of the suggested model remain as an important future work.

XI. Acknowledgments

We thank Zahide Pamir, Pinar Boyaci and Omer Faruk Gulban for insightful discussions on fMRI data pre-processing and GLM analysis. This project is supported by TUBITAK Project, 112E315. Mete Ozay was supported by the European commission project PaCMan EU FP7-ICT, 600918.

References

[1] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, “Beyond mind-reading: multi-voxel pattern analysis of fmri data,” *TRENDS in Cognitive Sciences*, vol. 10, no. 9, pp. 424 – 430, 2006.

[2] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, vol. 293, no. 5539, pp. 2425 – 2429, 2001.

[3] Y. Kamitani and F. Tong, “Decoding the visual and subjective contents of the human brain,” *Nature Neuroscience*, vol. 8, no. 5, pp. 679 – 685, 2005.

[4] J.-D. Haynes and G. Rees, “Predicting the orientation of invisible stimuli from activity in human primary visual cortex,” *Nature Neuroscience*, vol. 8, no. 5, pp. 686 – 691, 2005.

[5] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben, “Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection,” *NeuroImage*, vol. 28, no. 3, pp. 663 – 668, 2005.

[6] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wand, M. Just, and S. Newman, “Learning to decode cognitive states from brain images,” *Machine Learning*, vol. 57, no. 1-2, pp. 145 – 175, 2004.

[7] S. M. Polyn, V. S. Natu, J. D. Cohen, and K. A. Norman, “Category-specific cortical activity precedes retrieval during memory search,” *Science*, vol. 310, no. 5756, pp. 1963 – 1966, 2005.

[8] C. Plumpton, L. Kuncheva, D. Linden, and S. Johnston, “On-line fmri data classification using linear and ensemble classifiers,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4312–4315.

[9] C. O. Plumpton, “Online semi-supervised ensemble updates for fmri data,” in *Proceedings of the First IAPR TC3 Conference on Partially Supervised Learning*, 2012, pp. 8–18.

[10] A. Gramfort, G. Varoquaux, B. Thirion, and C. Pallier, “Decoding visual percepts induced by word reading with fmri,” in *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, 2012, pp. 13–16.

[11] L. Haufeld, R. Santoro, G. Valente, and E. Formisano, “Classification and visualization of multiclass fmri data using supervised self-organizing maps,” in *Proceedings of the 2012 Second International Workshop on Pattern Recognition in NeuroImaging*, 2012, pp. 65–68.

[12] M. Ozay, I. Oztekin, U. Oztekin, and F. T. Y. Vural, “Mesh learning for classifying cognitive processes,” *arXiv:1205.2382v2*, 2013.

[13] O. Firat, M. Ozay, I. Onal, lke Oztekin, and F. T. Y. Vural, “Functional mesh learning for pattern analysis of cognitive processes,” in *12th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC)*, 2013.

[14] I. Onal, M. Ozay, O. Firat, I. Oztekin, and F. T. Y. Vural, “Analyzing the information distribution in the fmri measurements by estimating the degree of locality,” in *Proceedings of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2013.

[15] ——, “An information theoretic approach to classify cognitive states using fmri,” in *13th IEEE International Conference on BioInformatics and BioEngineering (BIBE)*, 2013.

[16] C. Baldasano, M. C. Iordan, D. M. Beck, and L. Fei-Fei, “Discovering voxel-level functional connectivity between cortical regions,” in *Machine Learning and Interpretation in NeuroImaging Workshop, NIPS*, 2012.

[17] A. Zalesky, L. Cocchi, A. Fornito, M. M. Murray, and E. Bullmore, "Connectivity differences in brain networks," *NeuroImage*, vol. 60, no. 2, pp. 1055 – 1062, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811912000857

[18] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, 1973, pp. 267 – 281.

[19] G. E. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461 – 464, 1978.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465 – 471, 1978.

[21] I. Oztekin, C. E. Curtis, and B.McElree, "The medial temporal lobe and the left inferior prefrontal cortex jointly support interference resolution in verbal working memory." *Journal of Cognitive Neuroscience*, vol. 21, no. 10, pp. 1967–79, 2009.

[22] I. Oztekin and D. Badre, "Distributed patterns of brain activity that lead to forgetting." *Frontiers in human neuroscience*, vol. 5, p. 86, 2011.

[23] P. P. Vaidyanathan, *The Theory of Linear Prediction*. Morgan and Claypool Publishers, 2008.

[24] M. Xia, J. Wang, and Y. He, "Brainnet viewer: A network visualization tool for human brain connectomics." *PLoS ONE*, vol. 8, no. 7, 2013. [Online]. Available: http://www.nitrc.org/projects/bnv/