Fast Feature Extraction from Large Scale Connectome Data Sets Using Zero Crossing Counts Over Graphs

Panagiotis C. Petrantonakis
Information Technologies Institute
Centre for Research and Technology-Hellas (CERTH)
Thessaloniki, Greece
ppetrant@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology-Hellas (CERTH)
Thessaloniki, Greece
ikom@iti.gr

***Abstract*—Large scale connectome data sets based on sophisticated fMRI recordings are becoming prominent in the advent of big-data-in-neuroscience era. In this work, we present a fast and efficient connectome analysis method for feature extraction using randomly generated data over structural connectomes. The proposed approach is based on recursive filtering of the graph data and zero-crossing counting over the connectomes (graphs). The simplicity of the proposed method, namely simple Higher Order Crossings over graph sequence ($sHOC_g$), and its robust mathematical grounds, allow to discriminate subjects based on their structural wiring with high accuracy and dramatically faster estimation times (over 200 times faster) compared to state of the art graph kernel approaches.**

***Index Terms*—graph filtering, zero crossings, large scale connectome data set, graph classification**

I. Introduction

As large-scale neuroscience data collection initiatives grow, such as the Human Connectome Project (HCP) [1], the Adolescent Brain and Cognitive Development (ABCD) study [2], or the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3], there appears a great need for efficient and fast methods for the corresponding processing and analysis [4]. In particular, the contemporary extensive attempts for human brain mapping and the associated fMRI-based connectome data have put forward the need for graph analysis methods that are fast and efficient in various tasks.

In this work, we present the implementation of the simple Higher Order Crossings analysis over graphs ($sHOG_g$) method [5] for the feature extraction task while classifying large scale, braingraph data sets. More specifically, we use an augmented data set of human connectomes [6] and extract features using $sHOG_g$ and a state of the art Shortest Path kernel approach [7] to extract features. Two classification methods are used to classify connectomes originating from different subjects (classes), namely the k-Nearest Neighbor (kNN) [8], and the Binary Decision Tree (BDT) [9] classifiers. The speed and the performance of the two distinct approaches are investigated.

This work is supported by the project MINDSPACES that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825079.

The paper is structured as follows. The next section presents the materials and methods of this work. In particular the augmented connectome data set is briefly described whereas the proposed $sHOG_g$ approach is thoroughly presented. Moreover, the Shortest Path Kernel approach and the classifiers along with certain implementation issues are described. Section III presents the results of the application of the aforementioned methods and discusses certain aspects of the analysis. Finally section IV concludes the paper.

II. Materials and Methods

A. Human Connectome data set

The data set used in this work was presented in [6], [10] and is publicly available in https://braingraph.org/cms/download-pit-group-connectomes/. In the released contribution, a braingraph (connectome) set, computed from the 1200 Subjects Data Release of the Human Connectome Project [1]. The pre-processed 3T fusion data were used whereas the CMTK workflow [20] was utilized in the graph computation. For each subject, the segmentation and the parcellation steps were applied only once, whereas the probabilistic tractography part of the workflow was applied 10 times. The parcellation scheme was the Lausanne2008 atlas.

The connectome data set was augmented via a Newtonian Blurring method [6] as a modification of the Basic Averaging Strategy for braingraphs. In particular, for all subjects, the tractography step of the processing, which determine the axonal fibers, connecting the ROIs of the brain, is computed 10 times. Next, for each subject and each resolution (i.e., 83, 129, 234, 463, and 1015 nodes in the graph), the braingraph (connectome) of the subject is computed, and ten interim weights were assigned for each edge. The edges that appeared with 0 fibers in at least one of the 10 tractography runs, are deleted. For each subject and each resolution, 7 graphs from the 10 repeatedly computed graphs are chosen in every possible way (i.e., $\binom{10}{7}$ = 120). Subsequently, for each edge, the maximum and minimum edge-weights out of the 7 are

deleted, and the remaining five weights (number of fibers) are averaged (by simple arithmetic mean). This value is assigned to the edge as its final weight (these weights are also the ones used in this work). For more information about the data set construction the reader is encouraged to consult [6], [10].

B. Higher Order Crossings on Graphs

Higher order Crossings (HOC) analysis over graphs [5] is the sequential count of zero-crossings of a repeatedly filtered graph signal. HOC analysis of naive signals (not graph signals) have been shown to be an efficient method of feature extraction for classification tasks [11]–[13].

When a specific sequence of graph filters is applied to a particular signal, the corresponding sequence of zero-crossing counts is obtained, resulting in the so-called HOC_g sequence [5]. Different HOC_g sequences can be constructed by appropriate filter design.

Let $G = (\mathcal{V}, \mathbf{W})$ be a graph with $\mathcal{V} = (v_1, ..., v_n)$ being the set of nodes of the graph and $\mathbf{W}$ the weighted adjacency matrix of the graph. If the graph is an undirected graph then $W_{i,j} = W_{j,i}$. Given a graph G, we write a graph signal as a vector:

$$\mathbf{x} = [x_1, ..., x_n]^T \in \mathbb{R}^n \tag{1}$$

where each element x_i is associated with v_i of the graph $G = (\mathcal{V}, \mathbf{W})$.

To define the zero-crossings count for a graph signal, let the graph signal of (1) be a zero-mean graph signal. The associated clipped series are defined as:

$$y_i = \begin{cases} 1 & x_i \geq 0 \\ 0 & x_i < 0 \end{cases} \tag{2}$$

and the indicator function ζ_i for node i is simply the zero-crossings encountered in that specific node of the graph, i.e.,:

$$\zeta_i = \sum_{j \in \mathcal{N}_i} (y_i - y_j)^2 \tag{3}$$

where $\mathcal{N}_i$ is the set of nodes that are neighbors with the node i, i.e., $W_{i,j} > 0$. Thus, ζ_i is the number of zero-crossings that occur between node i and its neighbors. As a consequence, the total number of zero-crossings Z_g^0 of the initial graph signal in (1) within an undirected graph is:

$$Z_g^0 = \sum_{i=1}^{n} \zeta_i = \frac{tr(\mathbf{AY})}{2} \tag{4}$$

where $tr(.)$ is the trace of a matrix, $\mathbf{A}$ is the adjacency matrix of the graph G, i.e., $A_{i,j} = 1$ iff $W_{i,j} > 0$ and $A_{i,j} = 0$ iff $W_{i,j} = 0$, and $\mathbf{Y}$ is the matrix where $Y_{i,j} = (y_i - y_j)^2$ (see [5] for zero-crossing estimation on directed graphs).

Now let $\mathcal{L}_g$ be a high pass filter over the graph G. For a filtered graph signal $\hat{\mathbf{x}} = \mathcal{L}_g\{\mathbf{x}\}$ we can estimate Z_g^1, for $\mathcal{L}_g^2\{\mathbf{x}\}$ we estimate Z_g^2 and so on so forth in order to obtain the HOC_g sequence:

$$\mathbf{Z}_g = [Z_g^0, \ldots, Z_g^k] \tag{5}$$

where k is the maximum order of the sequence.

For the normalized HOC_g sequence, we divide the $\mathbf{Z}_g$ sequence, element-wise, by the maximum number of zero-crossings, Z_g^{max}, that can be detected for the complete graph G [5], thus,

$$Z_g^{max} = \begin{cases} \frac{n^2}{4}, & n \text{ is even} \\ \frac{(n-1)(n+1)}{4}, & n \text{ is odd.} \end{cases} \tag{6}$$

The normalized HOC_g sequence is estimated as:

$$\bar{\mathbf{Z}}_g = [\bar{Z}_g^0, \ldots, \bar{Z}_g^k] \tag{7}$$

where $\bar{Z}_g^i = \frac{Z_g^i}{Z_g^{max}}$. The normalized HOC_g sequence is the feature vector that is used for the classification of the brain connectomes in this work.

1) Filter Design: In this work we used graph filter of the form of polynomials on the weighted adjacency matrix $\mathbf{W}$, i.e.,:

$$h(\mathbf{W}) = h_0\mathbf{I} + h_1\mathbf{W} + ... + h_d\mathbf{W}^d, \tag{8}$$

where d is the degree of the polynomial. The output of the filter defined by (8) is the graph signal:

$$\hat{\mathbf{x}} = h(\mathbf{W})\mathbf{x}. \tag{9}$$

There are various approaches to design graph filters of the form (8) (see for example [5]). Nevertheless in this work we will adopt the simpler form, where $h_0 = 1$, $h_1 = -1$ and the rest parameters $h_i, i = 2, ..., d$ are zero. With this adoption we introduce the simple HOC_g sequence, i.e., $sHOC_g$ sequence which is merely the sequential weighted subtraction of each value in a vertex with its neighbors. This also resembles the simple HOC sequence introduced in [14] as the simplest form of high pass filter for ordinary signals (not graph signals).

C. Shortest path graph kernel

In order to compare the proposed approach for feature extraction based on $sHOC_g$ sequence we will also use the algorithm for Shortest Path graph kernel (SPK) by Borgwardt and Kriegel [7] that counts pairs of labeled nodes with identical shortest path length. The Matlab implementation that we used for this algorithm is publicly available at https://bsse.ethz.ch/mlcb/research/machine-learning/graph-kernels/graph-kernels.html. This approach was adopted here for comparison based on the fact that it is relatively fast in comparison with other graph kernels, e.g., Ramon and Gärtner [15], p-random walk [16] etc., with good performance in terms of graph classification task [17].

D. Classification

For the classification of the features extracted using the $sHOC_g$ and SPK approaches, two classifiers were used, the k-Nearest Neighbor (k-NN) classifier and binary decision tree classifier (BDT). For the k-NN, k was set to 3 neighbors whereas all parameters for BDT was the default parameters set by Matlab (all algorithms were written in Matlab 2019b).

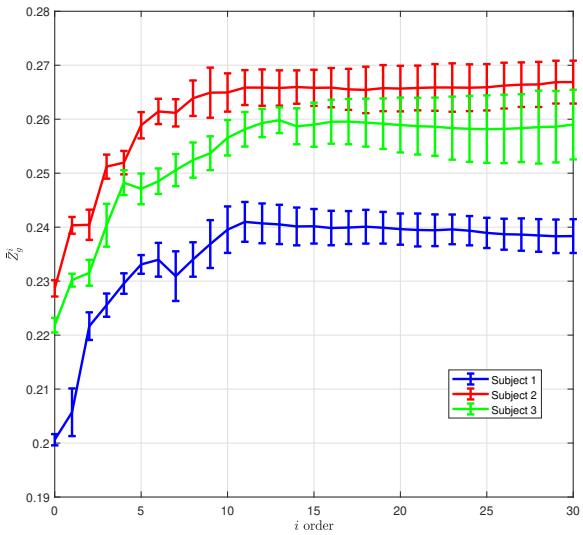

Fig. 1. Mean $\bar{Z}_g^i$ ($sHOC_g$) sequence and corresponding standard deviation with order of the sequence $k = 30$ for 3 different subjects (blue, red, and green lines) of the braingraph data set with resolution $|V| = 83$.

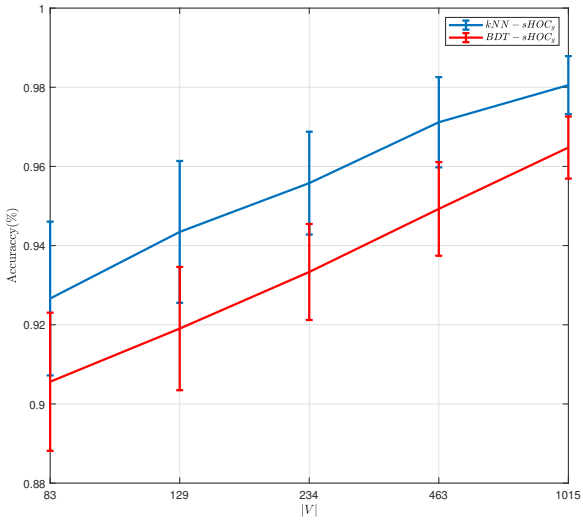

Fig. 2. Mean classification accuracy rates and corresponding standard deviation using 100 random graph signals and kNN and BDT classifiers for different braingraph resolutions, i.e., 83, 129, 234, 463, 1015.

The classification task corresponds to the classification of 50 different subjects (50 classes) in the connectome data set with 120 different connectomes per subject in a 10-fold cross-validation scheme.

III. Results and Discussion

For the evaluation of the proposed algorithm and the comparison with the SPK approach 50 subjects from the connectome data set were used. In particular, the first 50 subjects from the data set were used for the extraction of features using $sHOC_g$ and SPK for classification of the subjects based on 120 different braingraphs (connectomes) of each subject.

In order to extract the $sHOC_g$ sequence from the braingraphs, 100 random, uniformly distributed in the range $[-1\ 1]$ graph signals x were used and the performance of the classification of the $sHOC_g$ sequences based on these graph signals were estimated using the two classifiers. Fig. 1 represents the mean $sHOC_g$ sequence and corresponding standard deviation across 120 braingraphs with resolution $|V| = 83$ of three different subjects.

Fig. 2, shows the mean classification accuracy (in the range $[0\ 1]$, i.e., $0.5 - >50\%$ accuracy) and the associated standard deviation within the 100 different random graph signals. As was expected higher resolutions of the braingraphs lead to higher classification rates, from 92.66% ($|V| = 83$) to 98.05% ($|V| = 1015$) for the kNN classifier and from 90.56% ($|V| = 83$) to 96.48% ($|V| = 1015$) for the BDT classifier. Moreover, it is noteworthy that the standard deviation decreases as the resolution of the brain graph increases indicating a more robust classification with higher connectome resolutions i.e., from 1.94 ($|V| = 83$) to 0.73 ($|V| = 1015$) for the kNN classifier and from 1.74 ($|V| = 83$) to 0.78 ($|V| = 1015$) for the BDT classifier.

It is evident that certain graph signals can lead via the $sHOC_g$ sequence to advanced classification performance. This is confirmed for both classifiers. The association of the properties and characteristics of such graph signals with the properties of the graphs, i.e., subgraph structure, edges etc. remain to be investigated. Moreover, based on this investigation novel algorithms for such graph signal construction could be proposed and optimized in respect with fast and efficient graph classification. All this aspects will be the objective of future work.

In order to compare the proposed $sHOC_g$ sequence with a state-of-the-art approach we utilized the SPK method which was found to be among the fastest approaches and one of the more efficient within other graph kernel methods [17]. For the comparison, the graph signal with the best classification performance with $sHOC_g$ was used for each case.

Fig. 3 shows the accuracy rates (in the range $[0\ 1]$) using $sHOC_g$ (solid lines) and SPK (dashed lines) with kNN (blue lines) and BDT (red lines) across different resolutions of the braingraphs. kNN-SPK method exhibits the highest classification rate for all resolutions. Nevertheless, it is evident that classification rates with this method decrease as resolution increases. In particular, while for $|V| = 83$ accuracy for kNN-SPK is 99.97% it decreases to 99.82 for $|V| = 1015$. On the contrary, kNN-$sHOC_g$ approach leads to high classification accuracy rates as the resolution of the connectomes increases, i.e., from 97.13% for $|V| = 83$ it goes up to 99.73% for $|V| = 1015$, only 0.05% less than the kNN-SPK approach. Based on this aspect, it can be assumed that higher resolutions of the connectome would lead to even lower

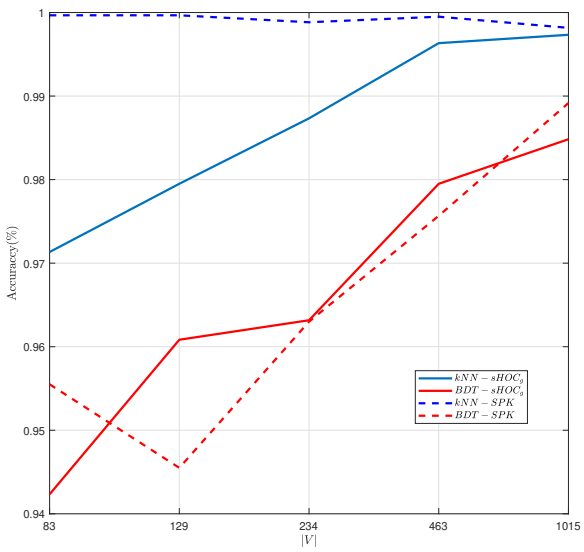

Fig. 3. Classification accuracy rates for different braingraph resolutions, i.e., 83, 129, 234, 463, 1015, using $sHOC_g$ (solid lines) and SPK (dashed lines) with kNN (blue lines) and BDT (red lines).

TABLE I
CLASSIFICATION ACCURACY FOR TWO SUBJECT GROUPS (GROUP 1: 1-50, GROUP 2:51:100 FROM THE DATA SET) USING THE GRAPH SIGNALS WITH THE BEST PERFORMANCE ON GROUP 1 WITH KNN CLASSIFIER

subject group/$\|V\|$	83	129	234	463	1015
group 1	97.13	97.95	98.73	99.63	99.73
group 2	98.43	94.97	97.78	98.72	99.23

accuracies for kNN-SPK and higher ones for kNN-$sHOC_g$ approach. Moreover, it should be stressed out that despite the fact that we chose the graph signals with the best classification accuracy for the $sHOC_g$ approach, these graph signals was the outcome of a mere random search and future optimization of the construction process of the graph signals will reveal new HOC_g construction methodology for even better performance. In addition, the performance of the BDT-$sHOC_g$ and BDT-SPK approaches seems to be approximately the same with the BDT-$sHOC_g$ approach outperforming the BDT-SPK one in three out of the five resolutions (see Fig. 3).

We also investigated if the graph signals that lead to high classification rates with the $sHOC_g$ method generalize well. In order to inspect such a behavior we estimated the classification performance of the proposed approach on another 50-subject group. In particular, whist the graph signals were tested on the classification performance within the 50 first subjects of the connectome data set (group 1) we also estimated the performance on the classification of the subsequent 50 subjects of the data set (group 2) using the initial graph signals that led to the best and worst performance on group 1.

Tables I and II show the accuracies for group 1 and group 2 using the graph signals with the best and worst performance on

TABLE II
CLASSIFICATION ACCURACY FOR TWO SUBJECT GROUPS (GROUP 1: 1-50, GROUP 2:51:100 FROM THE DATA SET) USING THE GRAPH SIGNALS WITH THE WORST PERFORMANCE ON GROUP 1 WITH KNN CLASSIFIER

subject group/$\|V\|$	83	129	234	463	1015
group 1	87.25	89.57	92.98	94.28	96.07
group 2	90.28	91.37	92.75	96.77	97.53

TABLE III
TIME (IN SECONDS) NEEDED TO EXTRACT THE $sHOC_g$ AND SPK FEATURES FOR 120 GRAPHS OF SUBJECT 1.

Method/$\|V\|$	83	129	234	463	1015
$sHOC_g$	0.04	0.06	0.16	0.85	4.89
SPK	0.34	1.09	5.12	74.37	1119.68

group 1, respectively, for the kNN classifier. Both for the best and worst cases (Table I and II, respectively), classification accuracies follow an increasing trend as the resolution increases as was initially observed for group 1. Moreover, accuracies for group 2 seem to depend on the efficacy of the corresponding graph signal to classify subjects in group 1. In particular, graph signals that tend to perform better in group 1 perform also with high accuracies for group 2. On the other hand graph signals that perform poorly on the classification task in group 1, perform poorly in group 2 as well. From the above it is evident that graph signals that are used to extract $sHOC_g$ sequence are important for the efficient performance of the classifiers due to probably inherent properties of the method that favor the discrimination of the different classes. It is noteworthy, that in some cases the classification rates of group 2 are higher than the ones in group 1 enhancing the above mentioned argument about the discrimination capabilities of the $sHOC_g$ approach. Hence, the extension of the proposed approach to facilitate, e.g., fMRI-EEG data [18], [19] that would lead to structure-function coupled features instead of random ones would further advance the performance of the method. In general, replacing the random signals with real experimental data would probably mitigate the issue of defining optimum graph signals for classification. Finally, it should be stressed out that despite the fact that Tables I and II present only the results for kNN classifier, the performance for the two groups is similar when using the BDT classifier.

So far the classification accuracy rates show that depending on the classifier and the graph signals used for $sHOC_g$ extraction, the proposed approach exhibits comparable and in some cases better performance than the SPK approach. Due to the fact that contemporary connectome data sets include braingraphs with high resolution which constantly increases, it is imperative that the feature extraction methodologies to be not only efficient but also fast. In Table III the times needed (all times are in seconds, and the analysis was performed in a desktop with Intel i5-9600K processor with 6 cores and 3.7GHz speed with a 16GB RAM) to extract the feature vectors of the 120 brain graphs from subject 1 in the connectome

data set are indicatively shown (similar times are needed for all subjects examined in this work). It is obvious that $sHOC_g$ approach is much faster that the SPK one. It is noteworthy that the ratios of the times needed to extract the SPK and $sHOC_g$ feature vectors, i.e., time for SPK/time for $sHOC_g$, rapidly increase as the resolution of the connectome increases. In particular, while SPK is approximately 8.5 times slower than $sHOC_g$ for $|V| = 83$, this ratio increases to approximately 229 for $|V| = 1015$. This is indicative of the superiority of the $sHOC_g$ for the analysis of large scale connectome data sets as with comparable performance (e.g., 99.82% for kNN-SPK vs. 99.73% for kNN-$sHOC_g$) $sHOC_g$ needs dramatically less time to be estimated. Taking into account that $sHOC_g$ construction can be further optimized via targeted graph signal construction, as previously discussed, the proposed method holds great potential for an efficient and a really fast analysis method for large scale, high resolution braingraphs [20], [21].

IV. Conclusions

In this work the $sHOC_g$ sequence was introduced and used for the extraction of features from braingraphs (connectomes) for their efficient and fast classification. Compared with state-of-the-art SPK method, $sHOC_g$ exhibited comparable, and in some cases better, performance. On the other hand, the proposed approach needs dramatically less time to be estimated making it an ideal tool in the arsenal for braingraph analysis tools for large scale data sets. The computational superiority of the proposed approach pave the way of feature extraction of voxel-wise brain graphs of even hundreds of thousands of nodes. Future work will also deal with optimized $sHOC_g$ sequence construction that would lead to even superior braingraph classification performance.

References

[1] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[2] N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway *et al.*, "The conception of the abcd study: From substance use to a broad nih collaboration," *Developmental cognitive neuroscience*, vol. 32, pp. 4–7, 2018.

[3] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga *et al.*, "Alzheimer's disease neuroimaging initiative (adni): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[4] R. N. Boubela, K. Kalcher, W. Huf, C. Našel, and E. Moser, "Big data approaches for the analysis of large-scale fmri data using apache spark and gpu processing: a demonstration on resting-state fmri data from the human connectome project," *Frontiers in neuroscience*, vol. 9, p. 492, 2016.

[5] P. C. Petrantonakis, "Higher order crossings analysis of signals over graphs," *IEEE Signal Processing Letters*, vol. 28, pp. 837–841, 2021.

[6] L. Keresztes, E. Szogi, B. Varga, and V. Grolmusz, "Introducing and applying newtonian blurring: An augmented dataset of 126,000 human connectomes at braingraph. org," *arXiv preprint arXiv:2010.09568*, 2020.

[7] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs," in *Fifth IEEE international conference on data mining (ICDM'05)*. IEEE, 2005, pp. 8–pp.

[8] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[9] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[10] B. Varga and V. Grolmusz, "The braingraph. org database with more than 1000 robust human connectomes in five resolutions," *Cognitive Neurodynamics*, vol. 15, no. 5, pp. 915–919, 2021.

[11] P. C. Petrantonakis and L. J. Hadjileontiadis, "Adaptive emotional information retrieval from eeg signals in the time-frequency domain," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2604–2616, 2012.

[12] P. Dickstein, J. Spelt, and A. Sinclair, "Application of a higher order crossing feature to non-destructive evaluation: A sample demonstration of sensitivity to the condition of adhesive joints," *Ultrasonics*, vol. 29, no. 5, pp. 355–365, 1991.

[13] S. He and B. Kedem, "Higher order crossings spectral analysis of an almost periodic random sequence in noise," *IEEE transactions on information theory*, vol. 35, no. 2, pp. 360–370, 1989.

[14] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, 1986.

[15] T. Gärtner, "A survey of kernels for structured data," *SIGKDD Explor. Newsl.*, vol. 5, no. 1, p. 49–58, Jul. 2003.

[16] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized kernels between labeled graphs," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, p. 321–328.

[17] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels." *Journal of Machine Learning Research*, vol. 12, no. 9, 2011.

[18] H. Behjat and M. Larsson, "Spectral characterization of functional mri data on voxel-resolution cortical graphs," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 558–562.

[19] J. Rué-Queralt, K. Glomb, D. Pascucci, S. Tourbier, M. Carboni, S. Vulliémoz, G. Plomp, and P. Hagmann, "The connectome spectrum as a canonical basis for a sparse representation of fast brain activity," *NeuroImage*, vol. 244, p. 118611, 2021.

[20] D. Abramian, M. Larsson, A. Eklund, I. Aganj, C.-F. Westin, and H. Behjat, "Diffusion-informed spatial smoothing of fmri data in white matter using spectral graph filters," *Neuroimage*, vol. 237, p. 118095, 2021.

[21] Y. Tian, B. T. Yeo, V. Cropley, A. Zalesky *et al.*, "High-resolution connectomic fingerprints: Mapping neural identity and behavior," *NeuroImage*, vol. 229, p. 117695, 2021.