Unsupervised Learning of Electrocorticography Motifs with Binary Descriptors of Wavelet Features and Hierarchical Clustering

Tim Pluta, Roman Bernardo, Hae Won Shin, and D.R. Bernardo

Abstract— **We describe a novel method for data mining spectro-spatiotemporal network motifs from electrocorticographic (ECoG) data. The method utilizes wavelet feature extraction from ECoG data, generation of compact binary vectors from these features, and binary vector hierarchical clustering. The potential utility of this method in the discovery of recurring neural patterns is demonstrated in an example showing clustering of ictal and post-ictal gamma activity patterns. The method allows for the efficient and scalable retrieval and clustering of neural motifs occurring in massive amounts of neural data, such as in prolonged EEG/ECoG recordings and in brain computer interfaces.**

I. Introduction

There is a need for the development of neuroinformatics data mining and analysis techniques to enable the study of functional brain network across multiple spectro-spatiotemporal scales concurrently [1],[2]. In this paper, we propose a novel system for the de novo computational discovery of recurring brain network motifs that occur over varying spectro-spatiotemporal scales in neural data. Neural activity arising from complex brain networks can be characterized by their specific structural or functional connectivity patterns or network 'motifs'[3],[4]. Network 'motifs' have been utilized in various fields to describe of recurrent patterns in a specific network or across several networks [5]. In signal analysis, time-series motifs can be characterized as recurring, similar subseries in a time-series dataset [6-8]. Recently the study of network motifs has been applied to electrophysiologic time series in the study of neuronal functional connectivity networks [3], [9], [10] .

The main issues for analysis of large volumes of neural data are (a) representing high-dimensional spectro-spatiotemporal features in a compact, noise robust manner, (b) efficient comparison of features across all subsequences, (c) and determining which features are correlated together. To motivate the need for a computationally efficient approach to these issues, consider that the typical ECoG recording consists of 100 separate channels recording at 1600 Hz. This represents ~14 trillion data points per 24 hours. In an *n*-item analysis with n = 8 bytes, an exact pair-wise comparison of this sample would require 264 gigabytes per sample per day. The sheer volume of data necessitates scalable data mining approaches to address current and future problems in clinical and research electrophysiology. To address these issues, we introduce a system of mapping neural spectro-spatiotemporal features to binary vector fingerprints. The use of compact binary codes allows for the efficient mining of preserved motifs using self-similarity search and co-clustering algorithms.

Our main contribution is a system for electrophysiological pattern recognition analysis that is scalable to large quantities of high-dimensional electrophysiologic data. Automated discovery of brain network motifs will be advantageous for advancing the diagnosis and treatment of neurologic and psychiatric diseases.

We demonstrate the fingerprinting algorithm in the investigation of network motifs consisting of gamma-range neural activities, which have found growing importance in the study of functional connectivity and in abnormal brain connectivity. For example, abnormal gamma oscillations have been implicated in the development of schizophrenia, [11] epilepsy, and Alzheimer's disease [12]. Evidence of conserved network motifs yielding characteristic gamma oscillation 'fingerprints' have been described in motor-evoked MEG activity and as the neurophysiologic basis of the default mode network (DMN) in ECoG [13], [14]. The strategy we developed is based on: 1) generation of compact binary codes representing spectro-spatiotemporal wavelet features of gamma activity 2) Unsupervised learning using hierarchal clustering. The rationale and framework for our strategy is described below:

II. Materials and Methods

A. Data Collection and Analysis

We retrospectively collected the electrocorticography (ECoG) data of a patient with medically refractory partial epilepsy whose ECoG data were obtained through an Institutional Review Board approved protocol that retrospectively identified patients who had underwent intracranial subdural electrode placement as a part of epilepsy surgery evaluation in the UNC Hospitals Epilepsy Monitoring Unit Database between the dates of 9/1/2011 and 9/1/2013. Continuous long-term recordings were obtained from Grass Recorder (Grass Technologies, Warick RI) with a sampling rate of 800 Hz. We selected ECoG data from the right frontotemporal subdural electrode grid. The sample

T. Pluta was with North Carolina State University, Raleigh NC 27607 USA. (e-mail: tim.pluta@gmail.com)

R. Bernardo was with the Department of Mathematics, UNC, Chapel Hill, NC 27599 USA.

H.W. Shin is with the Department of Neurology, University of North Carolina School of Medicine, Chapel Hill, NC 27599 USA

D.R. Bernardo was with the Department of Neurology, University of North Carolina School of Medicine, Chapel Hill, NC 27599 USA (e-mail: dbernard@unch.unc.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

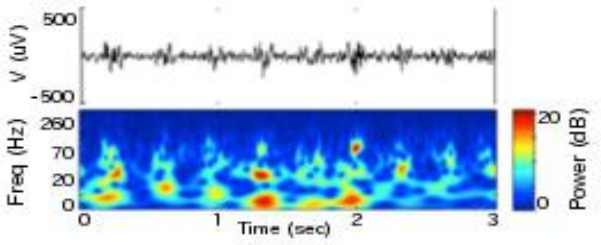

Figure 1. ECoG recording and corresponding Morlet CWT

used for analysis was a 1.5 hour long, 100 channel multi-electrode recording. The sampling rate was 800 Hz and the data was not filtered during pre-processing. All data analysis was performed in Matlab 2012b (The Mathworks, Natick, MA) using custom Matlab scripts on a computer with dual quad-core Intel i7 CPUs at 2.5 GHz and 16 Gb of RAM.

B. Spectral feature extraction

Initial dimensionality reduction of the ECoG dataset is performed by spectral feature extraction. Wavelet and spectral feature extraction have discriminative features which have been useful in developing brain computer interfaces (BCI) [15]. We convolved the ECoG with the complex Morlet continuous wavelet transform (CWT) as shown in Figure 1. To avoid edge effect from the complex Morlet CWT, we utilize an overlapping sliding window of the Morlet transform when designating peak points. We used a threshold using a 95% confidence level using a 1/f noise Markov Model of inherent brain noise, as described by Greenwood et al. as shown in Figure 2 [16]. Time-frequency peaks are then designated utilizing an intensity-weighted centroid blob detection algorithm. For study of gamma activity, detected blobs were included if they were greater than 50 ms, which is more than two low-gamma cycles.

C. Binary representation of ECoG data

For further dimensionality reduction, the spectral-spatial information can be represented in sparse two-dimensional binary codes. To the best of our knowledge, binary representation of EEG/ECOG data feature patterns has not previously been investigated. The representation of high dimensional data into binary codes has proven indispensable for the efficient indexing and machine learning in various applications such as chemoinformatics, bioinformatics, audio/video search, internet data mining, amongst others [17]. A common thread of different applications of binary encoding has been the "similarity-preserving" problem with the objective to map data containing similar features into similar-preserving binary codes. Similarly, the challenge in encoding neural data efficiently in compact, one-dimensional binary vectors is in the appropriate selection and preservation descriptive features of high dimensional neural data. Binary representation is suitable for study of gamma activities, which tend to consist of transient, episodic, bursts of activity. Discrete events of bursts gamma activity have been used previously in the study of co-occurring gamma activity and slow wave sleep [18]. The binary presence or absence of

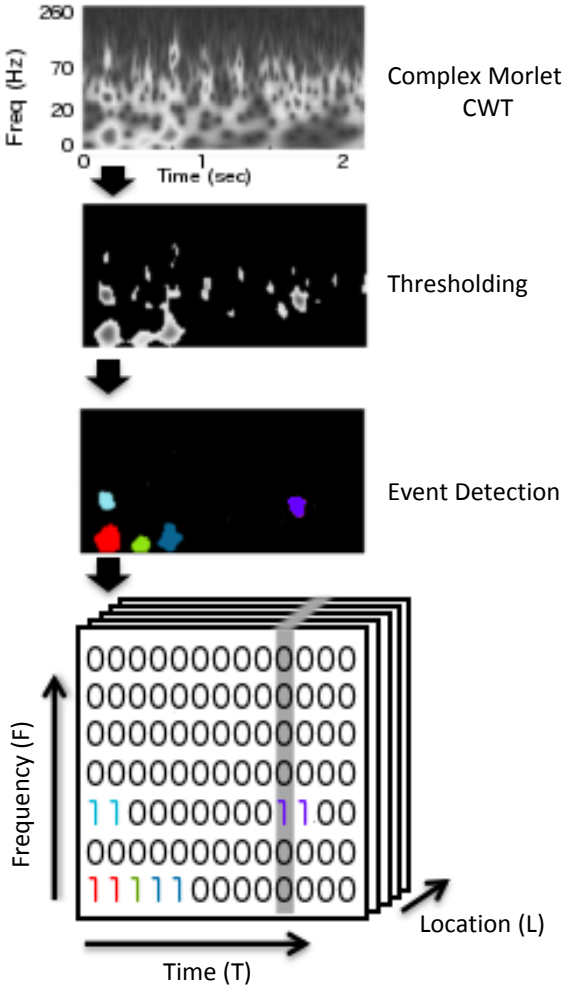

Figure 2. Schema for binary discretization of wavelet features

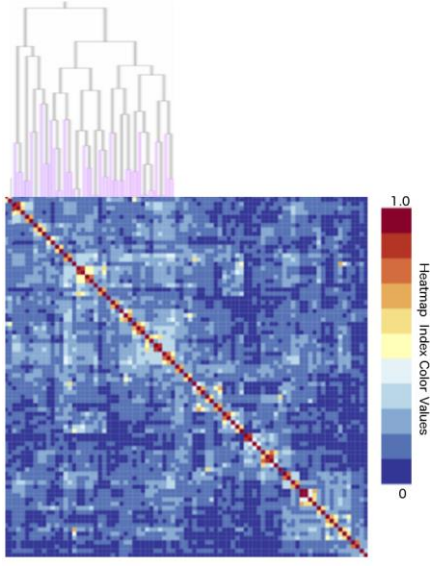

Figure 3. Sub-selection of heatmap dendrogram of clustered self-similarity matrix from yellow region of Figure 4. Higher values on the heatmap indicate higher Jaccard indices representative of more similarity.

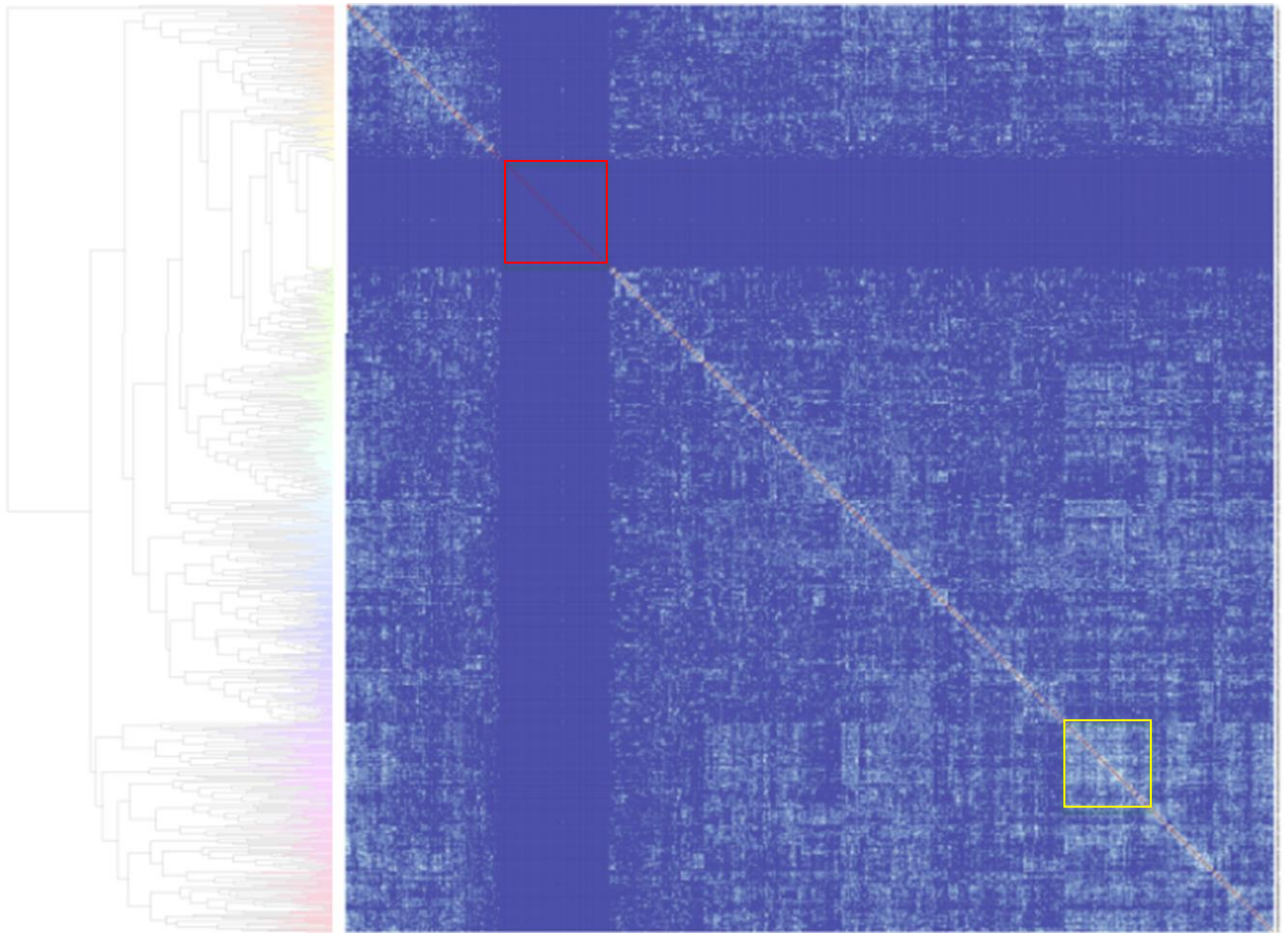

Figure 4. Heatmap dendrogram of clustered self-similarity matrix

gamma events has also been used in the study of co-activated structures as represented by interictal spike activity and in the study of functional connectivity as inferred from gamma events [19], [20].

D. Generation of binary vectors

We developed the use of a 3 dimensional binary matrix with 1 or 0 representing the presence or absence of a gamma event of at a frequency bin *F* at a given time bin *T* at a given spatial location *L* as shown in Figure 2. The spatial vector *L* represents intracranial electrode channel number and ranges from 1 to 100. The frequency vector *F*, ranges from 1 to 64 and represents equidistant log-scale frequency bins between 1 and 270 Hz. The spectral landmarks of neural activity recorded by ECoG are sampled into *N* time-bins with a finite time precision dT (i.e. $T = N * dT$). The choice of time-bin size is a compromise between i) combining separate peaks into one if a long-time bin is used (Type II error), ii) getting false positives co-occurrences of gamma activity if a long-time bin is used (Type I error). This is a similar problem to how the optimum time histogram bin size in the study of neuronal spike trains is best determined [21]. For the purposes of capturing co-activated gamma events, we selected a time-bin of 50 ms, the maximum period of duration of two individual, overlapping low gamma cycles (~40Hz). In comparison, 30 ms peri-event time bins were used recently to investigate connectivity estimated from gamma event maxima [20].

E. Generation of self-similarity matrices

To determine similarity between spectro-spatiotemporal features of shingles we used the Jaccard index, a statistic used to compare the similarity of sample sets. It is defined as the size of the intersection of the sets divided by the size of their union (1):

$$J(A,B) = |A \cap B| / |A \cup B| \quad (1)$$

The Jaccard index has been used in several applications of similarity search in a wide variety of domains including genome-scale clustering, web document similarity search, and molecular compound similarity search [22], [23]. Self-similarity matrices have previously been studied for demonstrating periodicities in time-series and complex biological networks [24]. To generate self-similarity matrices, we perform an exhaustive, all-pairs similarity join of each binary matrix *(L x F)* shingle at each *T*, shown in gray in Figure 3. Thus, the self-similarity matrix was defined as an *N x N* matrix of all pairwise Jaccard similarity coefficients between all *N*.

F. Cluster Analysis

We employ unsupervised cluster analysis of the ECoG data to accomplish motif discovery. Clustering has widespread use in bioinformatics and recently, in neuroinformatics clustering has been applied to the discovery of functional networks in fMRI [25]. Hierarchical clustering more recently has been used in automated analysis of interictal spike clustering and classification and in EEG artifact removal [25]. The use of hierarchical clustering of large-scale ECoG spectro-spatiotemporal features has not been previously been reported. Agglomerative hierarchical clustering analysis of the ECoG binary codes was performed using the *hclust* package in R using single linkage clustering and the Jaccard self-similarity matrix [26].

III. Results

A. Generation of Self Similarity Matrix

The generated self-similarity matrix for a 1.5 hr sample with time bin size of 50 ms measured 108,000 by 108,000 and was a sparse matrix populated by > 60% zeros. Greater than 95% of values had a Jaccard index of less than 0.1.

B. Visualization of Cluster Analysis

A 1-minute subsection of the self-similarity matrix containing a seizure was used for cluster analysis. We summarize the hierarchical clustering analysis of the self-similarity matrix in a cluster dendrogram heatmap matrix in Figures 3 and 4. Two visually apparent clusters indicated within the red and yellow areas were found to contain time periods containing post-ictal seizure state activity and seizure state activity, respectively. A zoomed-in view of the yellow area of Figure 4 is shown in Figure 3, with the representative, corresponding dendrogram for a selected cluster of the data.

IV. Conclusions

We have demonstrated a novel fingerprinting method of spectro-spatiotemporal motifs occurring in ECoG data. We illustrated how clustered self-similarity matrices may be potentially useful for identifying ictal and post-ictal states by allowing for the rapid visual analysis of recurrent motifs within a long ECoG sample.

References

[1] M. Helmstaedter and P. P. Mitra, "Computational methods and challenges for large-scale circuit mapping," *Current Opinion in Neurobiology*, vol. 22, no. 1, pp. 162–169, Feb. 2012.

[2] G. T. Einevoll, C. Kayser, N. K. Logothetis, and S. Panzeri, "Modelling and analysis of local field potentials for studying the function of cortical circuits," *Nat Rev Neurosci*, vol. 14, no. 11, pp. 770–785, Nov. 2013.

[3] O. Sporns and R. Kötter, "Motifs in Brain Networks," *PLoS Biol*, vol. 2, no. 11, p. e369, 2004.

[4] C. Echtermeyer, L. da Fontoura Costa, F. A. Rodrigues, and M. Kaiser, "Automatic Network Fingerprinting through Single-Node Motifs," *PLoS ONE*, vol. 6, no. 1, p. e15765, Jan. 2011.

[5] R. Milo, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.

[6] J. L. E. K. S. Lonardi and P. Patel, "Finding motifs in time series," *Proc. of the 2nd Workshop on Temporal Data Mining*, pp. 53-68. 2002.

[7] X. Du, R. Jin, L. Ding, V. E. Lee, and J. H. Thornton Jr, "Migration motif: a spatial-temporal pattern mining approach for financial markets," pp. 1135–1144, 2009.

[8] A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover, "Exact Discovery of Time Series Motifs.," pp. 473–484, Jan. 2009.

[9] M. Kaiser, "Integrating temporal and spatial scales: human structural network motifs across age and region of interest size," *Frontiers in neuroinformatics,* vol. 5 (2011). pp. 1–14, Jul. 2011.

[10] J. S. Anderson, M. A. Ferguson, M. Lopez-Larson, and D. Yurgelun-Todd, "Reproducibility of Single-Subject Functional Connectivity Measurements," *American Journal of Neuroradiology*, vol. 32, no. 3, pp. 548–555, Mar. 2011.

[11] P. J. Uhlhaas and W. Singer, "Abnormal neural oscillations and synchrony in schizophrenia," pp. 1–14, Jan. 2010.

[12] C. S. Herrmann and T. Demiralp, "Human EEG gamma oscillations in neuropsychiatric disorders," *Clinical Neurophysiology*, vol. 116, no. 12, pp. 2719–2733, Nov. 2005.

[13] A. L. Ko, F. Darvas, A. Poliakov, J. Ojemann, and L. B. Sorensen, "Quasi-periodic fluctuations in default mode network electrophysiology.," *Journal of Neuroscience*, vol. 31, no. 32, pp. 11728–11732, Aug. 2011.

[14] D. Cheyne, "MEG studies of motor cortex gamma oscillations: evidence for a gamma 'fingerprint' in the brain?" *Frontiers in human neuroscience* 7 (2013). pp. 1–7, Oct. 2013.

[15] P. Herman, G. Prasad, and T. M. McGinnity, "Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 16.4 (2008): 317-326.

[16] Erland S. Greenwood P.E., "Constructing 1/*f* noise from reversible Markov chains." *Phys. Rev. E*, 76, 2007.

[17] S.-S. Choi, S.-H. Cha, and C. C. Tappert, "A Survey of Binary Similarity and Distance Measures.," *Journal of Systemics, Cybernetics & Informatics*, vol. 8, no. 1, 2010.

[18] M. Le Van Quyen, J. Engel, et al. "Large-Scale Microelectrode Recordings of High-Frequency Gamma Oscillations in Human Cortex During Sleep" *Journal of Neuroscience*, vol. 30, no. 23, pp. 7770–7782, Jun. 2010.

[19] J. Bourien, F. Bartolomei, J. J. Bellanger, M. Gavaret, P. Chauvel, and F. Wendling, "A method to identify reproducible subsets of co-activated structures during interictal spikes.." *Clin Neurophys*, vol. 116, no. 2, pp. 443–455, Feb. 2005.

[20] F. Kheiri, A. Bragin, and J. Engel Jr, "Functional connectivity between brain areas.." *Journal of Neuroscience Methods*, vol. 214, no. 2, pp. 184–191, Apr. 2013.

[21] H. Shimazaki and S. Shinomoto, "A method for selecting the bin size of a time histogram.," *Neural Comput*, vol. 19, no. 6, pp. 1503–1527, Jun. 2007.

[22] A. Bender and R. C. Glen, "Molecular similarity: a key technique in molecular informatics," *Org. Biomol. Chem.*, vol. 2, no. 22, p. 3204, 2004.

[23] J. J. Jay, M. A. Langston, et al., "A systematic comparison of genome-scale clustering algorithms," *BMC Bioinformatics*, vol. 13, no. 10, p. S7, Jun. 2012.

[24] X. Xu, J. Zhang, and M. Small, "Superfamily phenomena and motifs of networks.," *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19601–19605, Dec. 2008.

[25] P. Wahlberg and G. Lantz, "Methods for robust clustering of epileptic EEG spikes," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 857–868, 2000.

[26] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics 5.3 (1996): 299-314*.