Improved clustering of spike patterns through video segmentation and motion analysis of micro Electrocorticographic data

Bugra Akyildiz*, Yilin Song*, Jonathan Viventi*, Yao Wang*
* Polytechnic Institute of New York University

***Abstract*—We have recently developed flexible, active, multiplexed recording devices for high resolution recording over large, clinically relevant areas in the brain. While this technology has enabled a much higher-resolution view of the electrical activity of the brain, the analytical methods to process, categorize and respond to the huge volumes of seizure data produced by these devices have not yet been developed. This paper exams a series of new measurements of vivo animal seizure recording based automatic spike segmentation, feature extraction, unsupervised clustering algorithm and quantitative evaluation for spike segmentation and clustering. We found significant improvement in spike classification accuracy by examining not only the pattern of individual channel but also the spatial temporal pattern variation over all adjacent channels. In this paper, we first applied advanced video analysis techniques (particularly region and motion analysis) for spike segmentation and feature extraction. Then we explored recent advances in machine learning for discovering useful features for clustering spike patterns and identifying natural clusters. After comparison base on evaluation matrices including Mutual Information Scores and intra cluster over inter cluster correlation, we found the best combination of feature set and clustering algorithm to be applying Dirichlet Mixture Model on raw video segmentation correlation matrix. The research is expected to yield new insights regarding how seizures initiate, progress and terminate, as well as subsequently significant improvement in seizure detection and prediction.**

I. INTRODUCTION

Currently many existing neurological data analysis rely on manual inspection. With the new technology we and others are developing to further enhance the spatial and temporal resolution, the data volume is infeasible for manual review. Further, manual inspection can miss subtle features that automated machine learning techniques can detect. There is an urgent need for efficient and sensitive automated methods that can analyze huge volumes of data.

One major application of our high resolution device is to record and examine brain signals in patients with epilepsy. Preliminary analysis of our high resolution data have discovered repetitive spatiotemporal microscale spike patterns that initiate and terminate seizures, not seen by standard electrodes. Understanding the ordering and relationships between these patterns is a key to develop better seizure detection and prediction techniques and ultimately better therapy.

In [2], one approach to recognize the pattern of epileptic seizure has been explored. In this early work, we identified spikes using a threshold-crossing based detector, extracted delay and power features from the detected spikes, and then performed clustering, allowing a gap-statistic to determine the number of clusters in the data. This revealed the presence of repetitive, stereotyped spatial patterns.

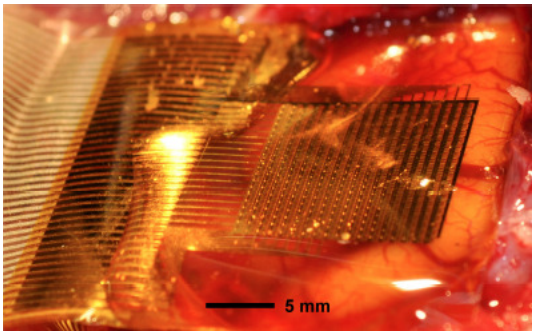


Fig. 1: Photograph of a 360 channel, high density neural electrode array used in a feline model of epilepsy. The electrode array is placed on the surface of visual cortex. The electrode size and spacing is 300 μm x 300 μm and 500 μm, respectively.

But simple scheme like thresholding, will generate disconnected parts or isolated scattered points even if the threshold is chosen properly. In this paper we treat the muti-channel signal as 3-D volume (successive 2D frames in time), and apply the 3D region growing technique to detect spike region. Region growing [6] is an effective technique to detect a connected (in 3D) region in which each pixel has an intensity that is noticeably higher (or lower) than pixels outside the region. We have also proposed a labeling algorithms which allows all the 3D segments have its unique identifier as their pixel value.

For each detected spike segment, we characterized its spatial-temporal pattern using some features. These features will be used for spike pattern clustering, as well as spike classification, and wavefront prediction. Intuitively, these features should characterize how the spike wavefront moves and be able to discriminate different moving patterns. Furthermore, we would like to use as few features as possible to achieve the same discrimination power. Last but not least, feature extraction computation complexity should be adjusted to suit for real time application.

The aim of this research task is to investigate whether all possible spike patterns form some natural clusters, so that patterns in each cluster can be approximated by the mean pattern of that cluster. We explore serveral well-known deterministic model including K-means algorithm, mean-shift algorithm. In addition to a deterministic classification model, we also explored probabilistic model as Gaussian Mixture Model (GMM) [8] for representing the presence of cluster within an overall samples. However a major drawback of a finite mixture model as GMM is that it will almost always use all components as much as they can, and hence will produce wildly different solutions for different numbers of components. Unlike finite models a Dirichlet Mixture Model [12] (DPM) solution wont change much with changes to the parameters, leading to more stability and less tuning. We will further demonstrate the difference between clustering method in Section IV & V.

The rest of this paper is organized as follows: In Section II, we present the algorithm we used for region growing & labeling. In Section III, several feature sets extracted from segmented regions are introduced. In Section IV, clustering models and methods are introduced. In Section V, several quantitative way to evaluate combinations of feature sets and classification models is introduced, as well as their each individual performance. Section VI concludes the paper.

II. SPATIAL-TEMPORAL SEGMENTATION

In this paper, a spike segment refers to a consecutive set of frames (each consists of multi-channel signal at the same time) in which one or several adjacent channels have high amplitudes. A simple scheme is by thresholding. This however will generate disconnected parts or isolated scattered points if the threshold is not chosen properly. In our preliminary work [4] [5] , we detect spike segment by first detecting the time when a single channel has a negative large signal exceeding a threshold (-0.5 mV), and then including all frames that is within a certain time window from this point both in the past (60 ms) and future (100 ms). Because each spike may last for somewhat different duration depending on its initial location and its moving speed, this simple scheme sometimes includes more than one spike or an incomplete spike in a detected segment. Given seizure video, we would like to develop an automatic scheme to detect wave pattern related to seizure by not only taking consideration of the amplitude of local signal but also the effect of its adjacent signal's amplitude. Meanwhile we want to distinguish spike segments that are not spatially connected by giving them different labels.

A. 3D region growing

In this paper, we treated the multi-channel signal as a 3-D volume (successive 2D frames in time), and apply the 3D region growing technique to detect the spike region. The algorithm of region growing first selects some pixels (called seeds) that have intensity values above a preset threshold. Each seed is an initial detected region. It then examines whether any immediate neighbors of a previously detected region boundary also have high intensity values, with a threshold that is determined based on the mean and standard deviation of the pixels inside the region. This process continues until no more pixels can be included. We have applied this algorithm to our ECoG data.

The algorithm works as in algorithm 1:

Algorithm 1 Region growing

$n \leftarrow 0$
if $\{P_{\vec{x}} > T, \vec{x_1} \in \mathbb{Z}^+_{18}, \vec{x_2} \in \mathbb{Z}^+_{20}, \vec{x_3} \in \mathbb{Z}^+_N\}$ **then**
  $\vec{x} \in \boldsymbol{S}_n$
end if
$\bar{P}_n \leftarrow E(P_{\boldsymbol{S}_n})$
$\bar{D}_n \leftarrow E[(P_{\boldsymbol{S}_n} - \bar{P}_n)^2]^{1/2}$
loop
  $\boldsymbol{S}_{n+1} \leftarrow \boldsymbol{S}_n$
  while $\vec{x} \in \boldsymbol{S}_n$ **do**
    if $P_{\vec{x} \pm \vec{e_i}} > \bar{P}_n - \alpha * D_n$ subject to $\vec{x} \pm \vec{e_i} \in$ dom **then**
      $\vec{x} \pm \vec{e_i} \in \boldsymbol{S}_{n-1}$
    end if
  end while
  $n \leftarrow n + 1$
  $\bar{P}_n \leftarrow E(P_{\boldsymbol{S}_n})$
  $\bar{D}_n \leftarrow E[(P_{\boldsymbol{S}_n} - \bar{P}_n)^2]^{1/2}$
  if $\boldsymbol{S}_n = \boldsymbol{S}_{n-1}$ **then**
    break loop
  end if
end loop

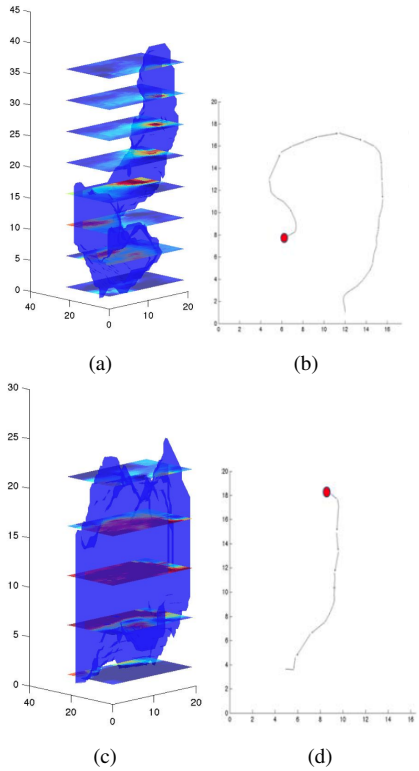
B. Labeling

After region growing, we get a binary representation of the raw signal with 1 representing foreground pixel related to seizure 0 representing background noise. In computer vision, labeled regions are good intermediated representation for regions that is helpful in further processing. There are multiple ways to representing regions, a simple approach is assign a unique integer to pixels belonging to a region.

After doing labeling, with the dataset we previously collected we are able to isolate 1262 disconnected segments after running region growing and labeling algorithm. However some of these segments only last up to 10ms, which not only have little benefit to study seizure pattern but also more prone to noise. Therefore we further determine to set a time constraint for the segmentations. By only selecting segments last longer than 54 ms, 508 segments remained fit for further study. Fig. 2 has shown two segments after region growing and labeling.

III. FEATURE EXTRACTION

For each detected spike segment, we would like to characterize its spatial-temporal pattern using some features. These features will be used for spike pattern clustering, to be discussed in Section V. Here we describe some of the features we have tried out. Intuitively, these features should characterize how the spike wavefront moves and be able to discriminate



(a) (b) (c) (d)

Fig. 2: Examples of spike segmentation. Subfigure a,b are for a spike with a spiral motion: Subfigure a shows the segmented volume in blue overlaid with the HR-ECoG signals captured at different times, with vertical axis corresponding to frame number; Subfigure b shows the trajectory of the centroid of the segmented region in successive frames, with the red dot indicating the beginning position; Subfigure c,d are for a spike with a vertical motion.

different moving patterns. Furthermore, we would like to use as few features as possible to achieve the same discrimination correctness. Eventually, the goodness of a feature set can only be judged by the classification and prediction accuracy it leads to. Therefore, feature extraction and evaluation is necessarily intertwined with feature based clustering and classification, and these techniques need to be jointly optimized in an iterative/multipass process.

A. Wavefront Trajectory

One way to characterize the motion pattern of a spike is to describe the trajectory of its wavefront over the duration of the spike segment, which is simply obtained by connecting the wavefront positions in successive frames. In Fig. 3 (c), we show trajectories representing the centroid of each cluster after running DPM (to be discussed in Section IV). For a spike segment consisting of $N_1N_2N_3$ samples (from N_3 frames of $N_1 * N_2$ channels), the trajectory can be captured by $2N_3$ features (including the wavefront 2D position at each frame). We further specified the area of detected region and the RMS signal power at each frame, with additional $2N_3$ features. These features together succinctly capture how the wavefront moves in time and how do the total number of active channels and their average energy change in time.

One problem with this representation is that N_3, the number of frames in each spike segment, and consequently the number of features for each segment is not a constant, which can lead to difficulty in feature-based clustering and classification. To get around this problem, we first pre-cluster the trajectory base on the length of the spike segments. For each segments within each cluster, resample each one to maintain the same feature dimension. An alternative approach is to calculate the correlation with each pair of trajectory, with shifting the shorter trajectory measuring the maximum correlation for the overlapping region with the longer one. Without losing generality, assume trajectories $X \in R^{2n}, Y \in R^{2m}, n \geq m$ Correlation of X, Y is measured under following term:

$$\begin{aligned} &\underset{i}{\text{minimize}}\ \text{Cov}(X_{(i:i+m)}, Y) \\ &\text{subject to } i \in \mathbb{Z}^{+}_{n-m} \end{aligned}$$

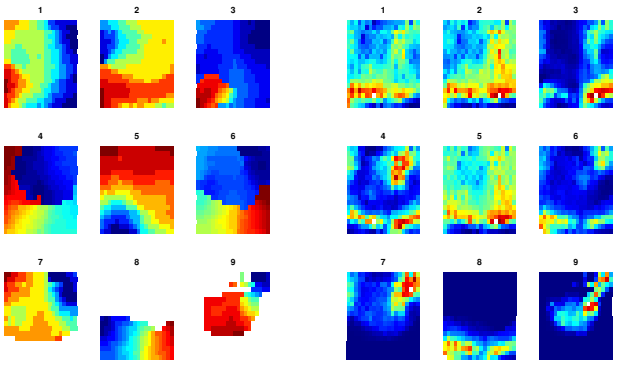
For each of the 508 segments, we measure the correlation with all the segments including itself and use this new trajectory correlation map as a new feature for further clustering.

B. Delay and Energy Map

Another way to characterize the spike motion is by computing the delays of individual channels with respect to a reference channel. In our prior work [2] , for each spike segment we compute a delay map, both have the same dimension as the 2D sensor array, and each element indicates the delay of the signal at that sensor with respect to the average of the signals or the strongest channel from all sensors over the time duration of the spike segment. As can be seen in Fig.3 (a), such a delay map is an effective and efficient way to characterize the spike motion. Unlike the trajectory representation which only characterizes the motion of the spike wavefront, the delay maps provide additional region shape information, by using $N_1 * N_2$ features. Furthermore, a summation of absolute signal value is calculated as energy map (Fig.3 b).

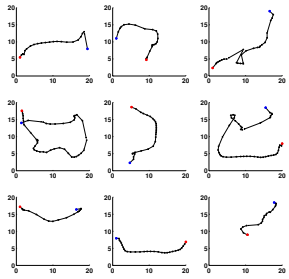
C. Segmented video signal correlation Map

The features listed above all serves the same purpose of dimension reduction which is necessary for real time classification. They are all partial representation of the true signal, by running whatever clustering algorithm on those feature would not give us a quantitative evaluation of correctness. To evaluate the correctness of classification we have to look back to the segmented region to evaluate the similarity between each two pairs. To tackle different length of segmentations and segmentations existing within the same time period but not spatially connected, we introduce M_X, which servers as a binary mask for the each specific segmentations X. Without losing generality, assume raw video signal $X \in R^{N_1 * N_2 * n}, Y \in R^{N_1 * N_2 * m}, n \geq m$ Correlation between two region are calculated as such:

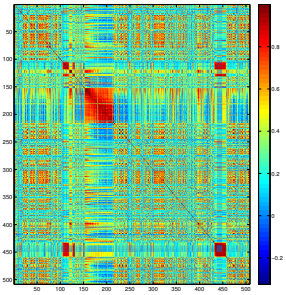


(a) delay map

(b) energy map



(c) trajectory



(d) correlation map

Fig. 3: Examples of extracted features. Subfigure a,b,c are delay, energy, trajectory representation of 9 spike segments closest to the cluster centroid after running DPM.
1) In a,b, blue indicates signal started earlier while red indicates late. White region in 7-9 shows it's not included in the spike.
2) In c, spike starting point is marked with blue, whereas ending point is marked with red. Each black dot represents the energy weighted centroid of one specific frame.
3) Subfigure d is the correlation map of all 508 segments. A similar coloumn vector indicates two segmentation belonging to the same cluster. For example similar column vectors between 150 to 220 could be interpreted as they form a natural cluster.

$$
\begin{aligned}
&\underset{i}{\text{minimize}}\ \text{Cov}(\hat{X}_{(,i:i+m)}, \hat{Y}) \\
&\text{subject to } i \in \mathbb{Z}^{+}_{n-m} \\
&\hat{X} = reshape(M_x \cdot X, N_1 * N_2, n) \\
&\hat{Y} = reshape(M_y \cdot Y, N_1 * N_2, m)
\end{aligned}
$$

Correlation of each pair of the raw video segments are calculated to form a raw signal correlation map. The raw signal correlation map would also serve as a quantitative reference for feature and clustering algorithm selection.

IV. Unsupervised Clustering of Individual Spike Pattern

Before running any clustering algorithm, we pre-processed the data to normalize its mean and variance to maintain consistancy across features. For each feature set discussed earlier or several sets combined, we tried to reduce the feature dimension in an unsupervised way to remove the intrinsic redundancy within the raw features by using principle component analysis (PCA). PCA is a mathematical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. Thus by appling an orthogonal transform, original $x^{(i)} \in \mathbb{R}^n$ feature renders a new vector $y^{(i)} \in \mathbb{R}^k$ with a lower dimension k, uncorrelated representation of $x^{(i)}$. We iteratively include transformed feature space with largest variance until $> 99\%$ of data variance was retained.

A. Deterministic Model

1) K-means algorithm: K-means[7] is one of the simplest and most frequently unsupervised learning algorithm which is mainly used for clustering purpose. It is used for partitioning N-dimensional observations into K disjoint observations by minimizing sum of least square criterion.

2) Mean Shift Algorithm: Mean Shift is an adaptive nonparametric iterative unsupervised learning algorithm [13], which unlike K-Means clustering algorithm does not require a prior knowledge of how many clusters are present in the observation space. It seeks mode to estimate clusters in the observation space. As K-means algorithm, Mean Sift Algorithm uses an iterative procedure to find the clusters in observation space, the different is Mean Shift removes one cluster once it converged to the center of the mass and run for another cluster until all the samples have labels.

B. Mixture model

In contrast to fitting a deterministic model for each of the feature set as k-means, an alternative one is mixture model approach. A mixture model is a mixture distribution over probability distribution of observations in the observation space [3]. A finite mixture model can be formulated as such:

$$p(x|\phi, \pi) = \sum_{k=1}^{K} (\pi_k p(x|\phi_k))$$

where $\pi = (\pi_1, \ldots, \pi_k)$, $\phi = (\phi_1, \ldots, \phi_k)$ are the parameters. One of the most common mixture model is Gaussian Mixture Model (GMM) where the $\phi_k = (\mu_k, \sum_k)$ and conditional density $p(x|\phi_k)$ is a Gaussian density with mean μ_k and covariance matrix $\sum_k$. Another mixture model is Dirichlet Process Mixture Model(DPM). The Dirichlet process is a prior probability distribution on clusterings with an infinite, unbounded, number of partitions.

1) Gaussian Mixture Model: A gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of gaussian distribution with unknown mean μ_k, covariance matrix $\sum_k$ and mixing coefficients π_k. Estimating μ_k and covariance matrix $\sum_k$ is to find maximum likelihood solutions for given dataset provided number of cluster k. An elegant and powerful tool is call *expectation-maximization* [10] algorithm or EM algorithm. It is the fastest algorithm for learning mixture models and as the algorithm maximizes only the likelihood it

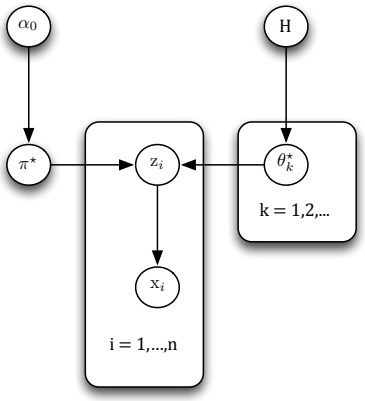


Fig. 4: Dirichlet Process Mixture Model

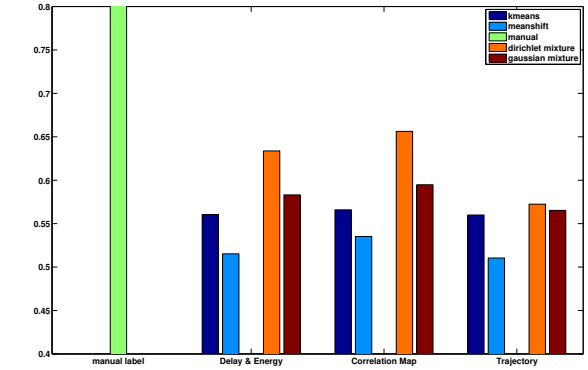


Fig. 5: Mutual Information Score of feature comparison and clustering method comparison. DPM outperforms other clustering method and correlation map is the best feature set. Inspect of rescaling the plots, Manual label actually have the score of 1.

will not bias the means towards zero or bias the cluster sizes to have specific structures that might or might not apply. That being said estimating number of cluster k is crucial for this algorithm as it will always use all the component it has access to. The Bayesian information criterion [11] (BIC) criterion can be used to select the number or components in a GMM in an efficient way. In theory, it recover the true number of components only in the asymptotic regime (i.e. if much data is available).

2) Dirichlet Process Mixture: An alternative approach to GMM is DPM, which has a variant of Gaussian mixture model with a variable number of clusters using the Dirichlet Process. A Dirichlet Process is a process that is a probability distribution over probability measures, which are random distributions[12]. Unlike finite models, which will almost always use all components as much as they can, and hence will produce wildly different solutions for different numbers of components, the Dirichlet process solution won't change much with changes to the parameters, leading to more stability and less tuning. The general struction of DPM is illustrated in Fig.4. With its parameter listed as following:

$$\pi^\star|\alpha \sim \text{GEM}(\alpha)$$
$$\theta_k^\star|H \sim H$$
$$z_i|\pi^\star \sim \text{Discrete}(\pi^\star)$$
$$x_i|z_i, \theta_{zi}^\star \sim \text{F}(\theta_{zi}^\star)$$

α is the concentration parameter that determines the likelihood of assigning a new cluster, and it is the only parameter in the algorithm. Distribution of π_k follows the Griffiths-Engen-McCloskey(GEM) distribution. H is the base distribution. To avoid sigularity of covarince matrix induced by insufficiently many points per mixture, H is selected to be gaussian distribution with diagnoal covarince matrix.

V. Evaluation Matric and Results

To evaluate all the feature sets and clustering method, we develop two kind of senarios. One is proved with ground truth, the other is without ground truth.

A. Mutual Information Score

We look at raw video signal and manually label each segmantation as ground truth. And use Mutual Information Score as evaluation of similarity between ground truth and target cluster. Mutual Information of two random variables is a quantity that measures the mutual dependence of two random variables by taking account of their common information. Assume two label assignments, U and V there entropy is the amount of uncertainty for a partition set, defined by:

$$H(U) = \sum_{i=1}^{|U|} P(i)log(P(i))$$
$$H(V) = \sum_{j=1}^{|V|} P'(j)log(P'(j))$$

where $P(i) = |U_i|/N$ is the probability that an object picked at random from U falls into class U_i, similarly $P(j) = |V_j|/N$ is the probability that an object picked at random from V falls into class V_j And mutual information between U and V is calculated by:

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)log(\frac{P(i,j)}{P(i)P'(j)})$$

where $P(i,j) = |U_i \cap V_j|/N$ is the probability that an object picked at random falls in both classes U_i and V_j. The normalized mutual information is defined as:

$$NMI(U,V) = \frac{MI(U,V)}{\sqrt{H(U)H(V)}}$$

As we can see from Fig.5, the descending order in terms of performance is correlation map, delay & energy map, trajectory. And the clustering method in a performance descending order is: DPM, GMM, K-means, Mean shift.

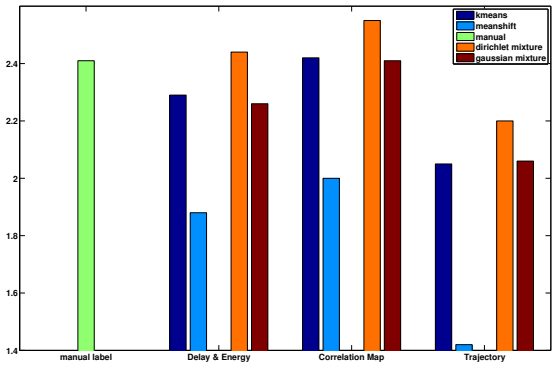


Fig. 6: Intra cluster over inter cluster correlation. Similar in Fig.5, intra over inter score renders a similar result in terms of best feature set and best clustering algorithm. And it also shows the human labeling is also reasonably high to be considered as a ground truth in the previous section.

B. Intra over Inter cluster corrlation score

Unfortunately, in the field of seizure prediction, HR-ECoG has not yet been used as a standard procedure. Hence we don't have an expert in the field to claim the label is correct, to demonstrate DPM and Correlation Map is still the best approach we use intra cluster over inter cluster raw video correlation as evaluation metric. Heuristically, a high intra over inter score indicates high consistancy within cluster and different result across clusters. The defision of intra, inter cluster correlation is denoted as:

$$\text{intra} = \frac{\sum_{i\in\mathbb{Z}_K^+} \sum_{\substack{a,b\in M^i \\ a\neq b}} \text{Cov}(M_a^i, M_b^i)}{\sum_{i\in\mathbb{Z}_K^+} \binom{N_i}{2} - N_i}$$

$$\text{inter} = \frac{\sum_{\substack{i,j\in\mathbb{Z}_K^+ \\ i\neq j}} \sum_{\substack{a\in M^i \\ b\in M^j}} \text{Cov}(M_a^i, M_b^j)}{\sum_{i,j\in\mathbb{Z}_K^+} N_i * N_j}$$

As can be seen in Fig.6, the results of best clustering algorithm and feature sets are the same.

VI. CONCLUSION

HR-ECoG has tremendous potential for many research and clinical application. In this paper, we have combined exciting new ideas in video analysis and machine learning with traditional signal processing method to look as these new datasets in novel ways. We have developed efficient methods for identifying and localizing the spatial and temporal extent of interictal and ictal spikes and detecting the spike wavefront, through advanced region and motion analysis. This new robust approach overcomes falsely connected and sperated spike segmentaion in simple scheme like thresholding and fix-window detection.

We also come up with multiple feature sets to represent the raw spike segmentation, in terms of offline clustering performance the best feature sets would be correlation map between different spikes. However in real time application, when a new spike segmentation comes in, we want to classify it into one existing cluster by computing the correlation between the cluster centroid. Assume each segmentation have similar length, the complexity for correlation map is $\sim O(kn_1n_2n_3)$ k being number of current clusters centroids. Whereas for delay map, correlation for each n_1n_2 channels is $O(n_3)$. Then use this n_1n_2 delay vector to compute euclidean distance with each cluster centroids whose complexity would be relatively trivial. And the total classification complexity for delay map would be $O(n_1n_2n_3)$, which is an enormous reduction of correlation map if the cluster number is high. And trajectory would serve as a within spike wavefront predition method. Hence at this point it's too early to tell which feature set is best for classification and prediction but all feature sets would have different roles in future seizure prediction.

In terms of clustering method, DPM is undoubtably the best clustering algorithm in terms of correctness. It's robust for parameter tuning and only need an upper bound for number of clusters. Although the extra parameterization necessary for variational inference will make inference slower, the clustering algorithm is running offline so this drawback is relatively trivial.

REFERENCES

[1] J. Viventi, D.-H. Kim, L. Vigeland, E. S. Frechette, J. a Blanco, Y.-S. Kim, A. E. Avrin, V. R. Tiruvadi, S.- W. Hwang, A. C. Vanleer, D. F. Wulsin, K. Davis, C. E. Gelber, L. Palmer, J. Van der Spiegel, J. Wu, J. Xiao, Y. Huang, D. Contreras, J. a Rogers, and B. Litt, Flexible, foldable, actively multiplexed, highdensity electrode array for mapping brain activity in vivo., Nature neuroscience, vol. 14, no. 12, pp. 1599 605, Dec. 2011.

[2] A. C. Chamberlain, J. Viventi, J. A. Blanco, D. H. Kim, J. A. Rogers, and B. Litt, Millimeter-scale epileptiform spike patterns and their relationship to seizures, in Proceedins of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 761764.

[3] C.Bishop (2006). Pattern recognition and machine learning. New York: Springer.

[4] J. Viventi, D.-H. Kim, L. Vigeland, E.S. Frechette, J. A. Blanco, Y.-S. Kim, A.E. Avrin, V.R. Tiruvadi, S.-W. Hwang, A.C. Vanleer, D.F. Wulsin, K. Davis, C.E. Gelber, L. Palmer, J. Van der Spiegel, J. Wu, J. Xiao, Y. Huang, D. Contreras, J. A. Rogers, and B. Litt, Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo, Nature Neuroscience, Nov. 2011, pp. 1-9.

[5] A. C. Chamberlain, J. Viventi, J. A. Blanco, D. H. Kim, J. A. Rogers, and B. Litt, Millimeter-scale epileptiform spike patterns and their relationship to seizures, in Proceedins of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 761 -764.

[6] R. Adams and L. Bischof, Seeded region growing, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 16, no. 6, pp. 641647, 1994.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, New York, New York, USA: Springer- Verlag, 2001.

[8] Permuter, H.; Francos, J.; Jermyn, I.H. (2003). "Gaussian mixture models of texture and colour for image database retrieval". IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP '03)

[9] B. Lucas and T. Kanade, An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. Seventh Intl Joint Conf. Artificial Intelligence, pp. 674-679, Aug. 1981.

[10] Dempster, A.P.; Laird, N.M.; Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B 39 (1), 1977, pp. 138.

[11] Schwarz, Gideon E. "Estimating the dimension of a model". Annals of Statistics 6 (2), 1978, pp. 461464.

[12] David Blei, Michael Jordan."Variational Inference for Dirichlet Process Mixtures" Bayesian Analysis (2006). 1, Number 1, pp. 121144.
[13] K. Fukunaga; L.D. Hostetler (1975). The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE Transactions on Information Theory, 21, 32-40.