Temporal Pattern Localization using Mixed Integer Linear Programming

Rui Zhao
ECSE Dept., RPI
Troy, NY, USA
zhaor@rpi.edu

Gerwin Schalk
Wadsworth Center, NYS Dept. of Health
Albany, NY, USA
schalk@neurotechcenter.org

Qiang Ji
ECSE Dept., RPI
Troy, NY, USA
jiq@rpi.edu

***Abstract*—In this paper, we consider the problem of localizing the subsequence in time series which contains the dynamic pattern of interest. This is motivated by brain computer interface application where we need to analyze the dynamic pattern of brain signals in response to external stimulus. We treat the localization as a binary label assignment problem and formalize a mixed integer linear programming (MILP) problem. The optimal solution is obtained by minimizing a cost function associated with label assignment subject to empirical constraints induced by data acquisition process. We first experiment with synthetic data to evaluate the effectiveness of the proposed MILP formulation and achieve 10.8$\%$ improvement on F_1-score. We then experiment with electrocorticographic (ECoG) data for a classification task and achieve 8.8$\%$ improvement on accuracy using subsequences localized by our method compared to other methods.**

I. Introduction

Time series data provide important information for the analysis of system dynamics. However, time series data are often contaminated with irrelevant part introduced during experiment process. Therefore the exact location and duration of the signals of interest in time series are often unknown. The major cause of this issue is that the data collection process itself is often insensitive to the dynamic pattern change over time. To be inclusive, data collection is often conducted continuously in time regardless of the presence or absence of interested event. For example, surveillance camera record videos all the time without knowing the presence or absence of abnormal event. The uncertainty in localization of dynamic pattern can also be caused by the technical limitation of data collection process. For instance, in the application of biophysical signal analysis, the data collection process follows a pre-defined experiment protocol. However, the starting moment of data collection may not be coincide with the actual response of the subject due to the variation of response time. Similarly, the stop moment of data collection is not aligned well with the end of subject response. In order to analyze the dynamic pattern possessed by signals of interest, it is crucial to identify the signals segment first. In this work, we address the problem of localizing signals of interest from discrete-time series containing irrelevant segments for the purpose of dynamic pattern classification. The rest of this paper is organized as follows. In section II, we review techniques to identify signals from un-segmented data. In section III, we define the problem and list the assumptions we use. We formalize the optimization problem in section IV. The experiments on synthetic data and electrocorticographic (ECoG) data are discussed in section V. Finally we draw conclusion and point out future work.

II. Related work

Existing approaches for temporal pattern localization in time series data can be divided into several major categories. The first major category treats time series data as a random process whose data points are samples from probabilistic distribution. The signals of interest are found by detecting change-point in time series. To declare a change-point, methods evaluating the difference between two distributions are used. The two major modeling issues need to be addressed in this category of approach is 1) how to characterize the probabilistic distribution given samples and 2) how to determine whether two distributions are different. To characterize the distribution, various statistics have been proposed to describe the random process [1], [2], [3]. Nonparametric density estimation is also used to allow more flexibility in fitting data [4]. The methods to determine the difference between distributions include two-sample Kolmogorov-Smirnov test [5], likelihood ratio test [6], [3], similarity test [7], [8], and density ratio test [9], [10]. The second major category is to formalize the localization as a segmentation problem where an optimal segmentation can produce the minimum cost in terms of the pre-defined optimization objective. The segmented data can then be used for pattern recognition. Representative works include piece-wise linear model [11], max-margin segmentation [12], [13], [14], [15], and clustering [16], [17], [18]. Keogh *et al.* [19] provided a comprehensive review on existing algorithms on segmentation of time series represented by piece-wise linear function. The third major category directly models time series using probabilistic dynamic models and convert the localization problem into a model fitting and probabilistic inference problem such as Gaussian AR [20], GGM [21], HMM [22], HSMM [23], [24], SLDS [25], [26], [27], HCRF [28], Coupled HMM [29], Sequence Memorizer [30], *etc.* An evaluation metric is derived from the fitted model to determine whether a change point occurred. Existing approaches often require the availability of ground truth label of whether a data sample belongs to the signals of interest in order to train the identification model. We propose an approach that can be trained unsupervisely by exploiting the relationship among

labels. This strategy is useful when ground truth labels are difficult to obtain such as in brain signals analysis application.

III. Problem Statement

A. Definition and Goal

We now formally define the temporal pattern localization problem as follows. Suppose we have a multi-dimensional time series $\mathbf{x}^L = \{x_0, ..., x_{L-1}\},\ x_i \in \mathbb{R}^d,\ \forall i = 0, ..., L-1$, where d is the dimension of each sample and L is the length of the sequence. Let $\mathbf{x}_n^l = \{x_n, ..., x_{n+l-1}\}$ $(0 \leq n \leq L-l,\ 1 \leq l \leq L)$ denote a subsequence consisting of l consecutive samples with the first sample be x_n. Let y_n^l be a binary variable indicating the label of subsequence $\mathbf{x}_n^l$. A subsequence $\mathbf{x}_n^l$ is called signal subsequence with $y_n^l = 1$ when it only contains samples from signals of interests. A subsequence that contains samples of irrelevant data is called irrelevant subsequence. The corresponding $y_n^l = -1$. The union of all signal subsequences is called signal segment. The remaining part is called irrelevant segment. For segmentation problem, we want to separate the signal segment from irrelevant segment. The overall goal is to learn a mapping function $f : \mathbb{R}^{dl} \to \mathbb{R}$. Then the label is decided as follows

$$y_n^l = \text{sign}(f(\mathbf{x}_n^l)) \tag{1}$$

Finally, we can identify the signal segment by taking the union of all the subsequences $\mathbf{x}_n^l$ with $y_n^l = 1$. An illustration of different components in a time series is shown in Fig. 1.

B. Assumptions

We assume each time series $\mathbf{x}^L$ contains one and only one signal segment of length $l' \geq L/2$. This majority assumption is reasonable in our application where signals of interest usually dominate during data collection process. We also assume the mapping function is a linear function of $\mathbf{x}_n^l$ which can be parametrized as follows

$$f(\mathbf{x}_n^l) = \mathbf{w}^T \mathbf{x}_n^l + b \tag{2}$$

Despite the use of linear model here, our formulation is general enough to allow the use of nonlinear model through kernel trick [31].

IV. Optimization Problem

In the sequel, we omit l in the superscript for compact notation. The value of l will be determined empirically during experiment satisfying $1 \leq l < L/2$. From Eq. (1), we have

$$y_n f(\mathbf{x}_n) > 0$$

In addition, we can scale the parameters $(\mathbf{w}, b)$ so that $\min(y_n f(\mathbf{x}_n)) = 1$, which yields

$$y_n f(\mathbf{x}_n) \geq 1$$

When the label y_n is available during training, one popular way of solving this problem is support vector machine (SVM) [32]. The parameters $(\mathbf{w}, b)$ are learned by solving the following optimization problem, which essentially maximizes the distance from decision boundary to the closest data point.

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ &\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1,\ n = 0, ..., L-l \end{aligned} \tag{3}$$

We can allow mis-classification by introducing positive slack variables.

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_n} \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{n=0}^{L-l} \xi_n \\ &\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \\ &\quad \xi_n \geq 0,\ n = 0, ..., L-l \end{aligned} \tag{4}$$

where γ is a constant weighting the penalty on mis-classification.

However, the challenge is that the ground truth label y_n is not available during training. Instead, we have some other information on y_n induced from the assumption that exactly one signal segment exists in a sequence. We reformulate the optimization problem as follows.

$$\begin{aligned} &\min_{y_n,\mathbf{w},b,\xi_n} \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_{n=0}^{L-l} \xi_n \\ &\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n,\ \forall n = 0, ..., L-l \\ &\quad \sum_{n=0}^{L-l} \frac{y_n + 1}{2} \geq 1 \\ &\quad \sum_{n=i}^{i+\lfloor \frac{L}{2} \rfloor - l} |y_{n+1} - y_n| \leq 2,\ \forall i = 0, ..., \lfloor \frac{L}{2} \rfloor - 1 \\ &\quad y_n \in \{-1, 1\},\ \xi_n \geq 0,\ \forall n = 0, ..., L-l \end{aligned} \tag{5}$$

where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. The second inequality constraint indicates the presence of at least one signal segment. The third set of inequality constraints requires the signal segment to have at least length $L/2$ since $|y_{n+1} - y_n| = 1$ indicates $\mathbf{x}_n$ is either the beginning or the end of signal segment.

We adopt the framework proposed by [33] to decompose Eq. (5) into two sub-problems. The overall optimization problem is solved iteratively by alternating between two sub-problems. For the first sub-problem, we assume y_n is known, then Eq. (5) reduces to Eq. (4), where we solve for $(\mathbf{w}, b)$. For the second sub-problem, we use $(\mathbf{w}, b)$ estimated in the first sub-problem and solve for y_n. We introduce new variables $u_n = \frac{y_n+1}{2},\ z_n = |u_{n+1} - u_n|$ and $a_n = \mathbf{w}^T\mathbf{x}_n + b$. The second

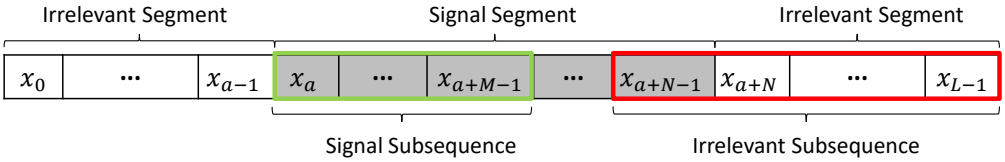

Fig. 1. Illustration of different parts of discrete-time sequence of length L. Each small rectangle represents a discrete-time sample. The length N signal segment $x_a, ..., x_{a+N-1}$ is shaded, where a is the starting index. The two white segments $x_0, ..., x_{a-1}$ and $x_{a+N}, ..., x_{L-1}$ are irrelevant segments. The subsequence in green bounding box is an example of length M signal subsequence and the subsequence in red bounding box is an example of irrelevant subsequence.

sub-problem is given by

$$\begin{aligned}
\min_{u_n, z_n, \xi_n} & \sum_{n=0}^{L-l} \xi_n \qquad (6)\\
\text{s.t. } & (2u_n - 1)a_n \geq 1 - \xi_n, \ \forall n = 0, ..., L-l \\
& \sum_{n=0}^{L-l} u_n \geq 1, \\
& z_n \geq u_{n+1} - u_n, \ \forall n = 0, ..., L-l-1 \\
& z_n \geq -u_{n+1} + u_n, \ \forall n = 0, ..., L-l-1 \\
& \sum_{n=i}^{i+\lfloor \frac{L}{2} \rfloor - l} z_n \leq 1, \ \forall i = 0, ..., \lfloor \frac{L}{2} \rfloor - 1 \\
& u_n \in \{0, 1\}, \ \xi_n \geq 0, \ \forall n = 0, ..., L-l
\end{aligned}$$

The second sub-problem Eq. (6) is a mixed integer linear programming problem. We solve the first sub-problem using `liblinear` [34]. We solve the second sub-problem by combining branch-and-bound and cutting plane methods [35]. The overall algorithm is summarized in Algorithm 1. Notice that the algorithm is described for one sequence. The generalization to multiple sequences is straightforward by adding the constraints associated with each sequence together.

Algorithm 1 MILP based temporal localization

Input: L: length of complete sequence, l: length of subsequence, S: number of iterations.

Output: Signal segment $\mathbf{x}^*$

1: Divide complete sequence into subsequence $\mathbf{x}_n$ and set initial value of $y_n, \forall n = 1, ..., L-l+1$
2: **for** $s = 1$ to S **do**
3: $\{\mathbf{w}, b\} \leftarrow$ solve Eq. (5) given $y_n, \mathbf{x}_n$
4: $\{u_n\} \leftarrow$ solve Eq. (6) given $\{\mathbf{w}, b\}, \mathbf{x}_n$
5: $y_n = 2u_n - 1$
6: **end for**
7: **return** $\mathbf{x}^* \leftarrow Union(\mathbf{x}_n), \forall y_n > 0$

V. Experiment

We perform experiments using both synthetic data and human brain signals.

TABLE I
Synthetic data segmentation results.

Method	F_1-score mean	F_1-score std.
MILP	0.9925	0.0073
Sign	0.8842	0.0276

A. Synthetic Data Experiment

The purpose of using synthetic data is to evaluate the quality of solution to Eq. (6) as we can compare the integer programming solution to the ground truth subsequence labels, which are known beforehand. We choose $L = 100, l' = 60, l = 20$ and randomly select a number α between 0 and $L - l'$. The subsequence labels $y_n, \ n = 0, ..., L-l$ are generated as

$$y_n = \begin{cases} 1, & \text{if } \alpha \leq n < \alpha + l' \\ -1, & \text{otherwise} \end{cases}$$

Then the corresponding sequence satisfies the assumptions we described in section III-B. To solve Eq. (6), we need to know the function value a_n for each subsequence $\mathbf{x}_n$. We generate a_n by drawing samples from normal distribution with mean depending on y_n.

$$a_n \sim \mathcal{N}(y_n, 1), \ y_n \in \{-1, 1\}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents normal distribution with mean μ and standard deviation σ.

We repeat the data generation process 10 times to get 10 sequences. Each sequence has an arbitrary starting point of signal segment. For each sequence, we solve Eq. (6) given a_n and then recover $y_n = 2u_n - 1$. For comparison, we use a naive approach (denoted as *Sign*), which decides label simply as $y_n = \text{sign}(a_n)$. We use F_1-score that is defined as follows to evaluate result.

$$F_1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The results in Table I shows about 11% absolute improvement on F_1-score with improved results close to 1. Therefore, the formulation of MILP is effective. By solving the problem with formulation specified by Eq. (6), we can identify the subsequence label correctly about 99% of the time.

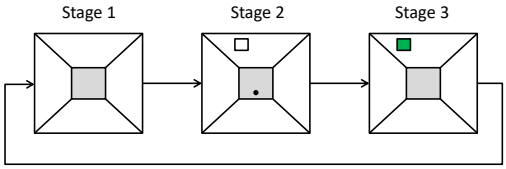

Fig. 2. Timeline of a trial. Stage 1, for 1 second, nothing appears on the screen depicted by the large square. Stage 2, a virtual target appears on the screen and a cursor appears on the screen immediately allowing subject to move the cursor controlled by a joystick. Either the target is successfully hit by the cursor or timeout after 2 seconds. Stage 3, the target will be highlighted for 1 second, indicating the end of trial. The next trial starts from stage 1 again.

B. ECoG Data Experiment

The ECoG data was recorded from a patient with electrodes placed subdurally on the surface of brain for solely clinical purpose of identifying epilepsy seizure foci prior to surgical resection. The subject had normal cognitive capability and was given informed consent. During the motion control experiment, the subject held a joystick to move a cursor appearing on the screen to hit a virtual target. The subject can move joystick to eight different directions. Each trial lasts 3-5 seconds with average hit time 1.4 second. The trial protocol is illustrated in Fig. 2. Multiple trials were recorded during experiment session. We select 20 sequences for each direction of movement as our dataset.

The ECoG signal is originally sampled at 1200Hz for each electrode. We follow the same process as in [29] to extract envelope features from γ band 70-170 Hz and downsample the signal to 400Hz. Fig. 3 shows two instances of filtered time series with color indicating signals from different electrodes. All sequences last 2 seconds, yielding $L = 800$ points for each sequence. The assumptions we described in section III-B is motivated by this application. Since we are only interested in the signal segment corresponding to the brain's response to external stimulus. In this case, the external stimulus is the visual cue of joystick movement. During the experiment, though we have accurate control of the visual cue onset, we do not know the accurate onset of brain signals due to different reaction time in different trials.

For real data we do not have ground truth label for subsequence. In order to evaluate the quality of identified signal segment. We perform a classification experiment where the classifier is learned using identified signal segments and original complete sequences. The overall experiment is divided into two phases namely the identification phase and the classification phase. In the identification phase, we perform segmentation for all the sequences simultaneously using Algorithm 1 by aggregating constraints from each sequence together. As an initialization, we choose $l = 200$ and the labels of subsequences are given by

$$y_n = \begin{cases} 1, & \text{if } n \geq 400 \\ -1, & \text{otherwise} \end{cases}$$

Therefore the presumed signal segment has length $l' = 600$. After the identification phase is finished. We obtain a set of signal segments from the original set of sequences. We found in experiment the localization results become stable after a few iterations. For comparison, we use other methods that can perform temporal clustering and localization including SC [36], ACA [16] and HMM+R [22]. For baseline, we use the initial assignment of signal segment. The second phase is classification where we split the dataset into two subsets. We train a HMM classifier using signal segment identified by each method of comparison from training subset and test the classification accuracy on testing subset. Since the same classifier is used, the final results only depend on the quality of localization.

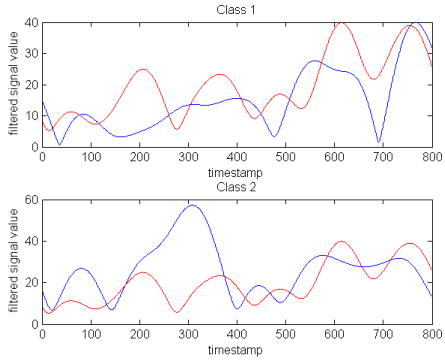

Fig. 3. Examples of different classes of multivariate ECoG time series. The raw signals are filtered for pattern analysis purpose. Different colors indicate data collected from different electrodes. Each time series contains 800 points which last 2 seconds. The starting point of the time series is aligned with the moment of visual cue displaying.

TABLE II
OVERALL ACCURACY OF ECOG DATA CLASSIFICATION RESULTS. CHANCE LEVEL = 12.5%

Method	Accuracy (%)
Initial	28.8
SC [36]	38.8
ACA [16]	41.8
HMM+R [22]	50.0
MILP	58.8

The overall accuracy of each method is shown in Table II. From the accuracy results we first see that all the methods exceed the chance-level. This indicates the presence of different dynamic patterns corresponding to different stimulus responses. Second, compared to baseline method, MILP improves the accuracy significantly by 30.0%. Compared to dynamic model based approach [22], MILP achieves 8.8% improvement. MILP also outperforms unsupervised clustering methods [16] and [36] by 17.0% and 20% respectively. These results demonstrate the effectiveness of the proposed approach.

VI. CONCLUSION

In this paper, we addressed the time series pattern localization problem motivated by brain signal analysis application. We formalize the localization as binary classification problem which decides whether a subsequence belongs to signal segment or not. We leverage on the empirical knowledge from data collection process to constrain that each sequence should

contain exactly one signal segment. To handle the unknown labels of subsequence during learning, we decompose the learning problem into two sub-problems by alternating the optimization between model parameters and label assignments. We solve the proposed mixed integer linear programming problem using branch-and-bound and cutting plane methods. We experiment this formulation on both synthetic data and real ECoG data. On synthetic data, we demonstrate the improvement of localization quality. On real data, we use classification as a surrogate metric to evaluate the quality of localization since the ground truth labels are not available. The improvement of classification accuracy shows the effectiveness of the localization method. For future work, we are interested in replacing the heuristics of using fixed subsequence length l and model it as a variable.

ACKNOWLEDGMENT

This work was supported by the NIH (EB00856, EB006356 and EB018783), the US Army Research Office (W911NF-08-1-0216, W911NF-12-1-0109, W911NF-14-1-0440) and Fondazione Neurone. We thank Dr. John E. Mitchell for the discussion on problem formulation.

REFERENCES

[1] F. Lombard, "Rank tests for changepoint problems," *Biometrika*, vol. 74, no. 3, pp. 615–624, 1987.

[2] M. Basseville, I. V. Nikiforov *et al.*, *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[3] M. Seck, I. Magrin-Chagnolleau, and F. Bimbot, "Experiments on speech tracking in audio documents using gaussian mixture modeling," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 601–604.

[4] E. Brodsky and B. S. Darkhovsky, *Nonparametric methods in change point problems*. Springer Science & Business Media, 2013, vol. 243.

[5] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[6] F. Gustafsson, "The marginalized likelihood ratio test for detecting abrupt changes," *IEEE Transactions on automatic control*, vol. 41, no. 1, pp. 66–78, 1996.

[7] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 180–191.

[8] H. Chen, N. Zhang *et al.*, "Graph-based change-point detection," *The Annals of Statistics*, vol. 43, no. 1, pp. 139–176, 2015.

[9] Y. Kawahara and M. Sugiyama, "Change-point detection in time-series data by direct density-ratio estimation," in *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 2009, pp. 389–400.

[10] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.

[11] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," *Data mining in time series databases*, vol. 57, pp. 1–22, 2004.

[12] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–1313.

[13] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, 2005.

[14] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–937.

[15] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3265–3272.

[16] F. Zhou, F. Torre, and J. K. Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–7.

[17] M. Hoai and F. De la Torre, "Maximum margin temporal clustering," in *Proceedings of international conference on artificial intelligence and statistics*, 2012, pp. 1–9.

[18] S. Tierney, J. Gao, and Y. Guo, "Subspace clustering for sequential data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1019–1026.

[19] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An online algorithm for segmenting time series," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001, pp. 289–296.

[20] K. Yamanishi and J.-i. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 676–681.

[21] X. Xuan and K. Murphy, "Modeling changing dependency structure in multivariate time series," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1055–1062.

[22] R. Zhao, G. Schalk, and Q. Ji, "Robust signal identification for dynamic pattern classification," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 3910–3915.

[23] X. Ge and P. Smyth, "Segmental semi-markov models for endpoint detection in plasma etching," *IEEE Transactions on Semiconductor Engineering*, vol. 259, pp. 201–209, 2001.

[24] Q. Shi, L. Wang, L. Cheng, and A. Smola, "Discriminative human action segmentation and recognition using semi-markov model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[25] Y. Kawahara, T. Yairi, and K. Machida, "Change-point detection in time-series data based on subspace identification," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 559–564.

[26] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, "Learning and inferring motion patterns using parametric segmental switching linear dynamic systems," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 103–124, 2008.

[27] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Nonparametric bayesian learning of switching linear dynamical systems," in *Advances in Neural Information Processing Systems*, 2009, pp. 457–464.

[28] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1250–1257.

[29] R. Zhao, G. Schalk, and Q. Ji, "Coupled hidden markov model for electrocorticographic signal classification," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 1858–1862.

[30] Y. Cheng, Q. Fan, S. Pankanti, and A. Choudhary, "Temporal sequence modeling for video event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2227–2234.

[31] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[32] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[33] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[35] J. E. Mitchell, "Branch-and-cut algorithms for combinatorial optimization problems," *Handbook of applied optimization*, pp. 65–77, 2002.

[36] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.