Dynamic Time-frequency Feature Extraction for Brain Activity Recognition

Yang Shi, *Student Member, IEEE,* Fangyu Li, Tianming Liu, *Senior Member, IEEE,*
Fred R. Beyette, *Member, IEEE,* and WenZhan Song, *Senior Member, IEEE*

***Abstract*— The biomedical signal classification accuracy on motor imagery is not always satisfactory, partially because not all the important features have been effectively extracted. This paper proposes an improved dynamic feature extraction approach based on a time-frequency representation and an optimal sequence similarity measurement. Since the wavelet packet decomposition (WPD) generates more detailed signal variation information and the dynamic time warping (DTW) helps optimally measure the sequence similarity, more important features are kept for classification. We apply the extracted features from our proposed method to Electroencephalogram (EEG) based motor imagery through the OpenBCI device and obtain higher classification accuracy. Compared with traditional feature extraction methods, there is a significant classification accuracy improvement from 83.53% to 90.89%. Our work demonstrates the importance of the advanced feature extraction in time series data analysis, e.g. biomedical signal.**

I. Introduction

Motor imagery helps disabled patients in moving wheelchairs [1], or do more complicated tasks while they have lost the abilities [2]. The electroencephalogram (EEG) is a typical non-invasive procedure for human brain disease diagnosis and disability assistance [1], [3]. Multiple kinds of EEG features have been selected for classification tasks, which mainly includes time domain features [3], frequency domain features [1], [2], [4], [5], and time-frequency domain features [6]–[10].

Power spectral density (PSD) has been used as a frequency domain feature extraction approach [4], [5] for the classification of motor imagery tasks based on EEG signals with an accuracy around 75%. The time-frequency domain feature applies to more aspects of classification tasks, including Epilepsy researches [6], [7], seizure detection [8], [9], and also motor imagery [10]. Time-frequency domain features express both time and frequency domain information, which obviously shows advantages over the features extracted from either domain [11].

To achieve improved classification results, we propose a dynamic time-frequency feature extraction approach applied to EEG signals for motor imagery, which combines a time-frequency representation method named wavelet packet decomposition (WPD) [10] and a dynamic temporal sequence similarity measurement named dynamic time warping (DTW) [12]. We perform quadratic discrimination analysis (QDA) [13] as the classification tool. In Section II, we introduce the basic conceptions about PSD, WPD, DTW and QDA, as well as the modifications we make. Section III describes the experiment dataset, settings, shown in Fig. 1, and classification results. In the end, we conclude the proposed approach and make discussions about the related applications.

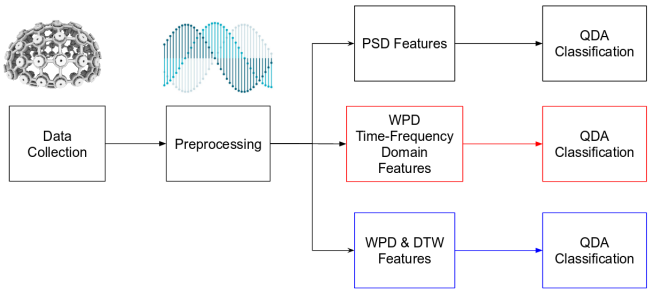

Fig. 1: System flow and comparison on feature extraction approaches.

II. Method

A. Power Spectral Density

As a common feature extracted from the frequency domain, power spectrum density (PSD) has been studied by biomedical research projects [4], [5]. We use the following equation to calculate the energy PSD_Δ under specific band Δ with the range $[l_\Delta, h_\Delta]$:

$$PSD_\Delta = \sum_{f=l_\Delta}^{h_\Delta} (|\mathscr{F}(f)|^2), \tag{1}$$

where, the spectrum at frequency f is denoted as $\mathscr{F}(f)$. We calculate PSD based on $\mathscr{F}(f)$ in a certain time window, given the lower frequency bound is l_Δ and the higher bound is h_Δ. We sum the power density from l_Δ to h_Δ to get the total power of the band. In each channel, according to [14], [15], we calculate PSD over four bands, respectively 4.5 to 8 Hz for θ wave, 8 to 12 Hz for α wave, 12 to 20 Hz for lower β wave, and 20 to 30 Hz for middle β wave. In Fig. 3, after the fast Fourier transform (FFT) plot, we calculate the PSD for the three events, noted as Event 1, 2, and 3, and plotted the stacked spectrum in black for comparison over

The research is partially supported by NSF-CNS- 1066391, NSF-CNS-0914371, NSF-CPS-1135814 and NSF CDI-1125165.

Yang Shi is with the Department of Computer Science and School of Electrical and Computer Engineering, University of Georgia, Athens, GA 30602, USA (e-mail: yang.atrue@uga.edu).

Fangyu Li, Fred R. Beyette and WenZhan Song are with the School of Electrical and Computer Engineering, College of Engineering, University of Georgia, Athens, GA 30602, USA (e-mail: fangyu.li@uga.edu, fred.beyette@uga.edu, wsong@uga.edu).

Tianming Liu is with the Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, University of Georgia, Athens, GA 30602, USA (e-mail: tliu@uga.edu).

the three events. Notice that though it can suppress noise, the "vanilla" stacking also smears certain features.

B. Wavelet Packet Decomposition

Either using time domain information or frequency domain information does not provide enough information in features, thus we use the time-frequency domain information as features for machine learning. Wavelet packet decomposition (WPD) is the tool for getting the features in time-frequency domain in this paper.

Assume that we have the input data from channel C, and the length of time window is T. We denote the training dataset as $X_0 = [\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, ..., \boldsymbol{x}_0^C]^T$. For every channel, we use the timing window as the input for WPD. The level of WPD is denoted as L, and the total number of band is 2^L.

The three WPD figures framed with red, green and blue in Fig. 3 are time-frequency representations from WPD with normalized magnitudes. In every sub-figure, the time domain signal is plotted on the top, and the spectrum is plotted on the right side. The figures show the signals vary in both time and frequency domain. WPD is known as an extension of discrete wavelet transform (DWT) [16]. It does not only decompose high frequency components of the signal for every level, but also decomposes low frequency components. And the signals are convolved by wavelets to get the components of the specific band.

C. Dynamic Time Warping

Algorithm 1 DTW

procedure DTW($\boldsymbol{a}, \boldsymbol{b}$)
  for $p = 1$ to n **do**
    $D(\boldsymbol{a}_p, 0) = \infty$
  end for
  for $q = 1$ to m **do**
    $D(0, \boldsymbol{b}_q) = \infty$
  end for
  $D(\boldsymbol{a}_0, \boldsymbol{b}_0) = 0$
  for $i = 1$ to n **do**
    for $j = 1$ to m **do**
      $d(i,j) = |\boldsymbol{a}_i - \boldsymbol{b}_j|$,
      $D(\boldsymbol{a}_i, \boldsymbol{b}_j) = d(i,j) + \min \begin{cases} D(\boldsymbol{a}_{i-1}, \boldsymbol{b}_j), \\ D(\boldsymbol{a}_i, \boldsymbol{b}_{j-1}), \\ D(\boldsymbol{a}_{i-1}, \boldsymbol{b}_{j-1}), \end{cases}$
    end for
  end for
  return $D(\boldsymbol{a}_n, \boldsymbol{b}_m)$
end procedure

The time domain information is not well addressed if we directly use the time-frequency domain features. Under spectral bands, the extracted features are sequences. The actual information of events within the sequences might be similar even when the frequency or offset of the sequences are different, and thus, we apply dynamic time warping (DTW) to calculate an optimal distance between two time series $\boldsymbol{a}$ and $\boldsymbol{b}$ with lengths of n and m respectively, as shown in Fig. 2 [17]. We calculate the optimal distance adaptively for getting maximal similarity between the two sequence data to address the time domain information better than traditional distance metrics. The way we calculate the distance is dynamic programming (DP), and each step of calculation in DP is shown in the Algorithm 1 [17].

Our next step is to use DTW as a constraint on features in QDA.

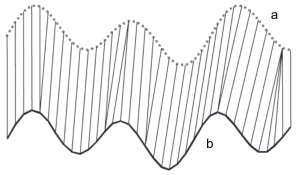

Fig. 2: Dynamic time warping of two signals, $\boldsymbol{a}$ and $\boldsymbol{b}$ [17].

D. QDA with DTW Constraint

Quadratic discrimination analysis (QDA) [13] is a classification model based on the estimation of probabilities of classes. We assume that the dataset from every label follows a Gaussian distribution, and the joint distribution is calculated as:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}, \quad (2)$$

where $f_k(\boldsymbol{x})$ is the probability density for class k at input $\boldsymbol{x}$, which is the features extracted from the training data. Then we denote the dimension as l, mean over dimensions as $\boldsymbol{\mu}_k$, covariance matrix as $\boldsymbol{\Sigma}_k$, and prior probability as π_k. Note that in order to classify the input into the label k, $P(G = k|\boldsymbol{X} = \boldsymbol{x})$ should be highest among all classes, where P is the probability when input $\boldsymbol{X}$ is $\boldsymbol{x}$, and the result of classification G is k. P is calculated via Bayes rule, and the detailed equation is:

$$P(G = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{i=1}^K f_i(\boldsymbol{x})\pi_i}. \quad (3)$$

Then the classification result based on the maximum of probability is calculated by the following equation:

$$\begin{aligned} \hat{G}(\boldsymbol{x}) &= arg\max_k Pr(G = k|\boldsymbol{X} = \boldsymbol{x}) \\ &= arg\min_k (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \\ &\quad + ln|\boldsymbol{\Sigma}_k| - 2ln\pi_k, \end{aligned} \quad (4)$$

which finishes the classification process as the output of the model.

In our approach, we use DTW features as constraints in the QDA model for classification. For every channel, we use the features under different bands. On time windows of every band, we apply DTW to dynamically calculate their optimal distances to the average of the specific band over whole training dataset as a feature. And after getting the optimized distance by DTW, we specify a weight α in the constraint of WPD features from DTW results. Then we evaluate the classification performance based on the constrained features. Mathematically, the new extracted features are denoted as

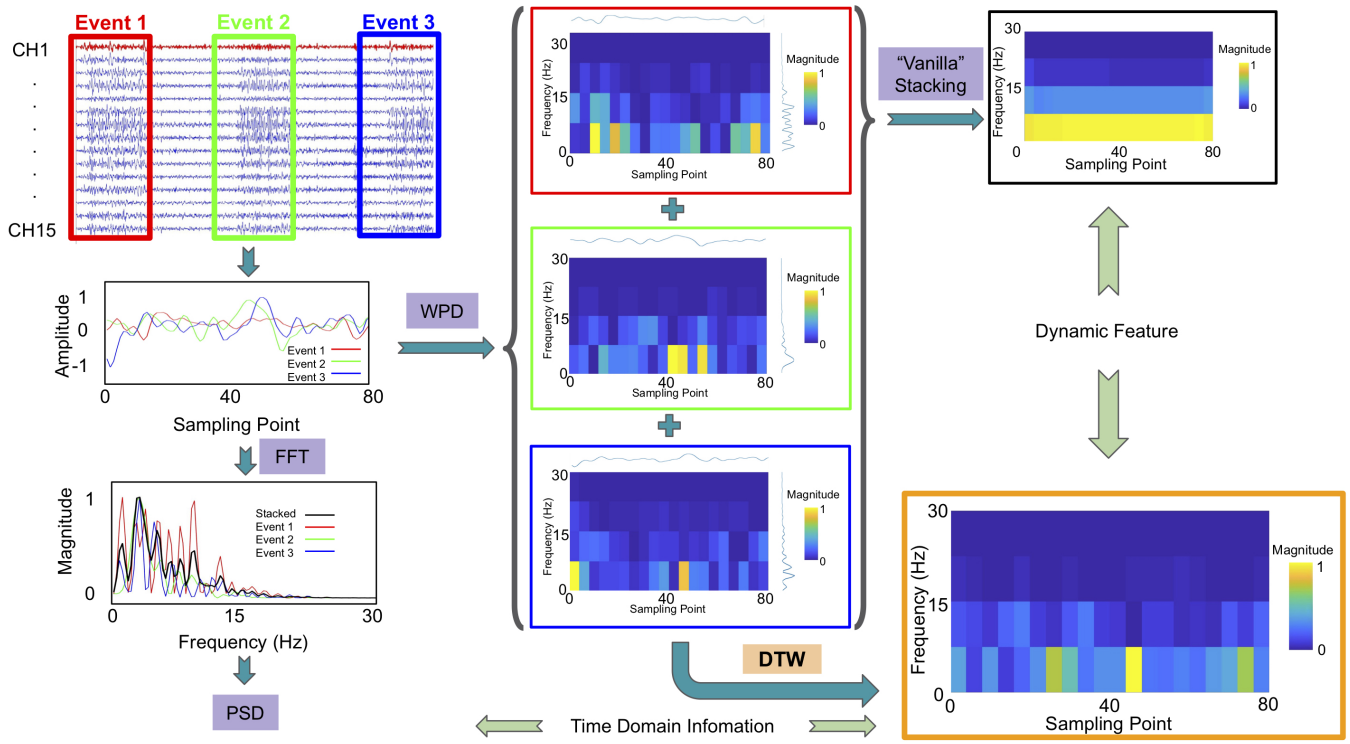

Fig. 3: Dynamic time-frequency feature extraction illustration. We use three consecutive events in Channel 1 (CH 1) as an example. In every time window associated with an event specified with red, green and blue frames, respectively, one sliding window with a length of 80 sample points is analyzed. We first plot the time series data in different colors. Through Fourier transform, we obtain the spectra of the events, based on which the black curve in the bottom left corner is the stacked spectrum. PSD then is calculated from the stacked spectrum. In another approach, WPD is applied to generate the time-frequency distributions, shown in the center. Keeping the temporal characteristics, we adopt a two level WPD with four spectral bands, 0 - 7.5 Hz, 7.5 - 15 Hz, 15 - 22.5 Hz and 22.5 - 30 Hz. If we use a "vanilla" stacking method for training WPD features, some important dynamic features will disappear, like the spectral information loss in the PSD calculation. To overcome the disadvantages of PSD in lacking temporal information as well as WPD in blurring dynamic patterns, we apply DTW in the feature extraction. Here, a DTW stacking example demonstrates the benefit from adaptive sequence similarity measurement. Compared with the two previous feature extraction approaches, the combination of WPD and DTW represents time-frequency patterns without losing important information during the stacking.

$\boldsymbol{D} = [D_1, D_2 ... D_{C2^L}]^T$ with level L, and the constraint is added as:

$$\boldsymbol{y} = [\boldsymbol{x}; \alpha \boldsymbol{D}]^T, \tag{5}$$

and with the constraint added, we calculate the classification result $\hat{G}_y(\boldsymbol{y})$ in the following equation:

$$\hat{G}_y(\boldsymbol{y}) = arg\min_k (\boldsymbol{y}-\boldsymbol{\mu}_{ky})^T \boldsymbol{\Sigma}_{ky}^{-1} (\boldsymbol{y}-\boldsymbol{\mu}_{ky}) + ln|\boldsymbol{\Sigma}_{ky}| - 2ln\pi_k, \tag{6}$$

where $\boldsymbol{\mu}_{ky}$ is the mean and $\boldsymbol{\Sigma}_{ky}$ is the covariance matrix for the constrained input denoted as $\boldsymbol{y}$.

III. Motor Imagery Application

A. Dataset and Preprocessing

We collected the data from OpenBCI 16 channel headsets [18], and conditioned them to exactly 250 Hz for the feature extraction methods. A band-pass filter between 3 and 30 Hz is then applied to suppress interferences. The data collection consists of 15 random trials of tasks imagining of moving left or right hands for continuous 4 seconds, and between each trial, the subject was asked to rest for 6 seconds. The subject is asked to perform the experiments in the common lab space with no other interference.

B. Feature Extraction Comparison

TABLE I: Confusion matrix based on PSD features.

$\hat{G}\backslash G$	Rest	Left	Right
Rest	88.02% (18487)	6.69% (1405)	5.29% (1111)
Left	15.23% (993)	73.94% (4819)	10.83% (706)
Right	11.48% (674)	10.43% (613)	78.09% (4589)

We evaluate the three approaches on the dataset for comparison, as shown in Fig. 1. The evaluations are based on 5-fold cross-validation. In the first experiment, we apply PSD features, and use four bands as specified in [14], [15]. We visualize the stacked features from the three experiments in Fig. 3. The classification accuracy based on PSD features is 83.53%, which is similar to the previous research results mentioned in [4]. The detail of confusion matrix can be found in Table I. The notation of the classification result is $\hat{G}$, and the ground truth is noted as G.

We further evaluate the feature extraction approach using WPD. The classification accuracy is 88.31%, which outperforms the performance of PSD features by 5%. Table II shows the details for this approach evaluation. The usage of time domain information from WPD is the reason of the improvement, compared with PSD approach.

TABLE II: Confusion matrix based on WPD features.

$\hat{G}\backslash G$	Rest	Left	Right
Rest	91.47% (19211)	4.81% (1009)	3.72% (781)
Left	11.18% (728)	82.16% (5355)	6.67% (435)
Right	10.97% (650)	5.19% (308)	83.84% (4968)

Then we use the DTW extracted features with a weight α as the constraint of QDA, the accuracy of classification result $\hat{G}_y$ improves further to 90.89%. Table III is the confusion matrix for this experiment and the accuracy is especially high for distinguishing left and right. The misclassification rate between them is about 1.5%. This improvement comes from the application of DTW, which makes additional usage of time information in actual motor imagery events like "left" and "right" in our experiments.

TABLE III: Confusion matrix based on constrained features in QDA.

$\hat{G}_y\backslash G$	Rest	Left	Right
Rest	92.24% (19372)	4.90% (1029)	2.87% (602)
Left	8.96% (584)	89.23% (5816)	1.81% (118)
Right	10.63% (630)	1.43% (85)	87.93% (5210)

IV. Discussions and Conclusions

Our experiments show that dynamic features extracted from the time-frequency domain bring considerable classification accuracy improvement for motor imagery tasks. The reason behind is not complicated, employing more valuable information is always beneficial. The main question asked in pattern recognition tasks remains what are the intrinsic features and how to extract them. In this study, WPD attains both time and frequency domain information, which brings more available features. The only adaptive operation we adopted is the dynamic calculation of the similarity of two sequences which minimizes the errors brought from the dislocations between two sequences. Though, we did not specify the actual bands we intend to use, the classification results are already better. To further improve the accuracy, one promising research topic would be isolate the effective band of every brain wave adaptively.

In this study, we first extended the frequency domain features into time-frequency domain features, which improved the classification accuracy by around 5%, and then use DTW constraint in QDA classification, which improved accuracy again by about 3%. Our proposed approach of dynamic time-frequency feature extraction produces high accuracy. Because our approach solves a general problem in biomedical signal classification, we expect its wide applications would impact a variety related tasks in areas such as brain decoding. By applying the dynamic time-frequency feature extraction, our algorithm reduces the computational cost compared to deep models, and thus is available to limited computation source devices for real applications.

References

[1] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[2] K. LaFleur, K. Cassady, A. Doud, K. Shades, E. Rogin, and B. He, "Quadcopter control in three-dimensional space using a noninvasive motor imagery-based braincomputer interface," *Journal of Neural Engineering*, vol. 10, no. 4, pp. 046 003+, Aug. 2013.

[3] V. Srinivasan, C. Eswaran, and a. Sriraam, "Artificial neural network based epileptic detection using Time-Domain and Frequency-Domain features," *Journal of Medical Systems*, vol. 29, no. 6, pp. 647–660, Dec. 2005.

[4] R. Chatterjee and T. Bandyopadhyay, "EEG based motor imagery classification using SVM and MLP," in *2016 2nd International Conference on Computational Intelligence and Networks (CINE)*. IEEE, Jan. 2016, pp. 84–89.

[5] A. Konar and S. Saha, "EEG-gesture based artificial limb movement for rehabilitative applications," in *Gesture Recognition*, ser. Studies in Computational Intelligence. Springer International Publishing, 2018, vol. 724, pp. 243–268.

[6] E. Bagheri, J. Jin, J. Dauwels, S. Cash, and M. B. Westover, "Fast and efficient rejection of background waveforms in interictal EEG," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016, pp. 744–748.

[7] M. Li, W. Chen, and T. Zhang, "Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble," *Biomedical Signal Processing and Control*, vol. 31, pp. 357–365, Jan. 2017.

[8] D. Chen, S. Wan, J. Xiang, and F. S. Bao, "A high-performance seizure detection algorithm based on discrete wavelet transform (DWT) and EEG," *PLOS ONE*, vol. 12, no. 3, p. e0173138, Mar. 2017.

[9] B. Boashash, H. Barki, and S. Ouelha, "Performance evaluation of time-frequency image feature sets for improved classification and analysis of non-stationary signals: Application to newborn EEG seizure detection," *Knowledge-Based Systems*, vol. 132, pp. 188–203, Sep. 2017.

[10] Y. Zhang, B. Liu, X. Ji, and D. Huang, "Classification of EEG signals based on autoregressive model and wavelet packet decomposition," *Neural Processing Letters*, vol. 45, no. 2, pp. 365–378, 2017.

[11] B. Nguyen, D. Nguyen, W. Ma, and D. Tran, "Wavelet transform and adaptive arithmetic coding techniques for EEG lossy compression," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, May 2017, pp. 3153–3160.

[12] A. Bajoulvand, R. Zargari Marandi, M. R. Daliri, and S. H. Sabzpoushan, "Analysis of folk music preference of people from different ethnic groups using kernel-based methods on EEG signals," *Applied Mathematics and Computation*, vol. 307, pp. 62–70, Aug. 2017.

[13] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, Mar. 2002.

[14] E. Başar, C. Başar-Eroglu, S. Karakaş, and M. Schürmann, "Gamma, alpha, delta, and theta oscillations govern cognitive processes," *International Journal of Psychophysiology*, vol. 39, no. 2-3, pp. 241–248, Jan. 2001.

[15] Y. Yasui, "A brainwave signal measurement and data processing technique for daily life applications," *Journal of PHYSIOLOGICAL ANTHROPOLOGY*, vol. 28, no. 3, pp. 145–150, 2009.

[16] Y. Kutlu and D. Kuntalp, "Feature extraction for ECG heartbeats using higher order statistics of WPD coefficients," *Computer Methods and Programs in Biomedicine*, vol. 105, no. 3, pp. 257–267, Mar. 2012.

[17] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, Mar. 2005.

[18] P. J. Durka, R. Kuś, J. Żygierewicz, M. Michalska, P. Milanowski, M. Łabcki, T. Spustek, D. Laszuk, A. Duszyk, and M. Kruszyński, "User-centered design of brain-computer interfaces: OpenBCI.pl and BCI appliance," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 60, no. 3, Jan. 2012.