Feature Extraction Method for EEG based Biometrics

Sukun Li\textsuperscript{a} and Sung-Hyuk Cha\textsuperscript{b}

Department of Computer Science, Pace University, New York, NY, 10038, USA

E-mail: \textsuperscript{a}sukun.li@pace.edu; \textsuperscript{b}scha@pace.edu

Abstract. Distinct feature extraction methods are simultaneously used to describe single channel Electroencephalography (EEG) based biometrics. This study proposes a new strategy to extract features from EEG signals. Based on the time and frequency information, the statistics features are obtained from the EEGs. For the dichotomize process, the support vector machine classifier is used with 10-cross-fold in this research. The main contribution of this paper is to propose a simple but effective single-channel EEG feature extraction method and consider feature selection to optimize classification efficiency. In the experiments, the EEG data is obtained from a human-computer interaction environment when the subjects are under the non-stationary states with different emotions. The results show that this proposed method achieves better classification performance on a single-channel EEG system than previous work.

1. Introduction

Electroencephalography (EEG) signals are used for user recognition and are considered to be biometric technology. The EEG-based biometrics have the advantages of continuous authentication and the robustness against spoofing attacks which could improve or complement the traditional biometrics. However, it still faces problems for enhancing the recognition accuracy and robustness on practical applications. The user recognition process mainly includes three steps: data pre-processing, feature extraction and classification. Extracting stable features from an acquired signal is a critical step to improve the biometric performance, and this paper focuses on the feature extraction of the high dimensional EEG data for biometric authentications.

Features are represented in terms of quantitative numerical values calculated using mathematical processing from EEG signals. When in purpose representative the biometric signals, the features analysis methods are mainly capturing linear or nonlinear changes in the time or frequency domain. At present, feature extraction methods for the EEG biometrics mainly include: Autoregressive (AR) Coefficients, Power Spectral Density (PSD), statistics features etc. In [1–4], the Yule-Walker or Burg’s method of AR was implemented as self-regressive of the signal in the time domain or in the frequency domain when concerning the power spectral density of each EEG signal. In [5–8], a high dimensional PSD with fast Fourier transform (FFT) was conducted as the stable features on either an EEG identification or authentication system. The statistical approach was suggested for representing the probability density of EEG signals to the practical analysis [9–13]. This type of approach comprises the statistical cumulants in either time or frequency domain. In [13], the statistical features applied as well for event-related potential on user identification study as the responses to visual stimuli. Study[11] used a 6-dimensional statistic feature vector is specified as the characteristics of the signal. The feature set provided the authentication performance of 80.60% for intra-class and inter-class classification with the Support Vector Machine (SVM) classifier.
This paper proposed a simple but useful feature extraction method for EEG biometric authentication system. The method is inspired by study [10]. We utilized the time series and the differences of the time series as the data sequences, and then extract the quantitative information and entropies of these sequences as features. The motivation for using these features is that they depict the characteristics of EEG signals in time and frequency domain which provides multidimensional information. We applied the dichotomized model in this research and the SVM, based on statistical learning theory, was used in this study. Finally, the correlation-based feature selection (CFS) algorithm was utilized with a greedy algorithm in order to find the best classification results according to the proposed method. This proposed method achieved the classification accuracy rate to 96.90% on the AF3 channel which enhances 12.75% than the previous study.

The remainder of this paper is divided into three sections: Section 2 describes the method for EEG feature extraction and provides details. Section 3 describes the experimental protocols, and finally, the paper concludes in Section 4.

2. Methods

In this paper, the EEG is modeled as a random signal with a stochastic process, and a distribution which characterizes this process was determined. The proposed approach includes two steps: 1) compute the amplitude distribution of the EEG data, 2) extract the structural features of the distribution. Assume \( S = (s_1, s_2, \ldots, s_n) \) is the vector that represented the EEG sequence for data points \( N = \{1,2,\ldots,n\} \). Then the distribution of this sequence is estimated as

\[
H(s|\xi) = H(s_1, s_2, \ldots, s_n|\xi)
\]

where the \( \xi \) are parameters which denote the specific shape of the distribution and provide quantitative information on various signal attributes.

A dynamic analysis method was applied to the pre-processed signal sequence in order to measure the change of differences in the time domain. The first difference and the second difference of the data sequence were computed. Hence, the \( H(s|\xi) \) represents the distribution of the original data sequence, the first difference of the data sequence, and the second difference of the data sequence, respectively. To represent the shape of the distributions, a multidimensional vector \( \xi \) has been extracted from the distributions. Let \( M \) be the order statistical moment of the distribution and is given by

\[
M^s_i = \begin{cases} 
\frac{1}{N} \sum (S - \mu)^2 & \text{for } i = 1 \\
\frac{1}{N} \sum \left( \frac{S - \mu}{\sigma} \right)^i & \text{for } i > 2 
\end{cases}
\]

where \( \sum \) is the expectation operation of data \( S \), \( \mu \) is the linear average of the amplitude, and \( \sigma \) is the standard deviation. The mode and median are applied to represent the histogram shape, and the \( z \)-normalization is utilized to describe the structural similarities of the data sequence in the vector \( \xi \).

Furthermore, Entropy is another useful quantitative information to recognize the chaos of the non-stationary EEG signals. We selected the Approximate Entropy (ApEn) and Sample Entropy (SampEn) to analyze the chaos in time domain, and the Shannon Spectral Entropy (ShanEn) is used to supplement the spectral information in frequency domain. To deal with the biometric authentication issue, a dichotomy transformation process is applied to measure the statistical inferable discrimination. Let \( f^n_d \) be the \( d^{th} \) feature vectors of \( n^{th} \) biometric data, while the feature extracted from the biometric dataset could be shown as \( \{f^n_1, f^n_2, \ldots, f^n_d\}, \{f^n_1, f^n_2^2, \ldots, f^n_d^2\}, \ldots, \{f^n_1^N, f^n_2^N, \ldots, f^n_d^N\} \). According to the dichotomized transformation algorithm [14–15], the extracted features are transformed from a feature space to a feature distance space by computing the feature distance of each pair. The distance measure in this algorithm depending on the type of feature [5], but here an absolute difference is used.

After calculating the distance of two features, a labeling process is applied to those distanced features as an inter-distance class or intra-distance class. Specifically, if the distance of EEG features
was extracted from the same person, we labeled it as intra-distance features; otherwise, we labeled them as inter-distance features. Those intra-distance features are close to each other and near the origin coordinate but the inter-distance features are sparsely distributed in the coordinate system of the feature distance domain. Finally, those features were classified via SVM classifier with 10-fold cross-validation to determining whether the EEG feature belongs to the authenticated person or not.

3. Experiment Results

3.1. Dataset and setup
The practical application we consider was an authentication system for EEG based biometrics. In order to prove the impact of the proposed method on classification results, we used the same time period of the signal data and classification model [11], which is to represent the emotional state EEG signals for analysis. The EEG data was collected from 32 healthy subjects. The data was collected using an EEG sensor from AF3 channel: Electrode location for collection of EEG data. The electrodes attached to subjects following the international 10–20 standard [16]. During data collection, the subjects were instructed to seat and watching multimedia and breathe normally. The EEG data was extracted from the raw signal with downsampling to 128Hz after applying a bandpass filter to restrict the available frequency range from 4 Hz to 45 Hz.

For the binary classification, the two categories of the distance classes corresponding to “1” as the same person, and “0” as the different persons respectively. According to the method as mentioned in Section 2, three feature set was prepared in our experiments. The size of each prepared feature set is 1000 which contains 500 intra-distance features and 500 inter-distance features.

3.2. Results and Discussion
As shown in Table 1, with a dichotomy transform model our method outperformed the latest reported results on the emotional EEG based biometric authentication system. As can be seen, the best classification accuracy for the EEG signal was 96.90% which is significantly improved than [11].

To observe the feature influence on the curve of dimensional for classification results, we evaluated the different feature combinations in the proposed method. The vector $\xi$ with features $M_1^X$, $\sqrt{M_2^X}$, $M_3^X$, $M_4^X$, $\frac{\mu}{\sigma}$, median, mode, ApEn, SamEn, ShanEn were examined on the original data sequence, the first difference of the data sequence, the second difference of the data sequence. The average classification accuracy rate for the feature sets and their combination with different entropies are shown in Table 2.

| Method      | Dimensionality | Feature Representation                  | Classifier | ACC (%) | FRR   | FAR   |
|-------------|----------------|----------------------------------------|------------|---------|-------|-------|
| [11]        | (6)            | Mean of Amplitude                      | SVM        | 84.15   | 22%   | 9%    |
| proposed    | (30)           | Distribution of Amplitude + Spectral Entropy | SVM        | 95.70   | 6%    | 2%    |
| proposed    | (23)           | Distribution of Amplitude + Spectral Entropy | SVM        | 96.90   | 4%    | 2%    |

As can be seen from the Table 2, only the entropy features for amplitude are insufficient to represent the EEG pattern, but when combined with the chaos attribute of the frequency, the classification performance can be improved. From the results in Table 2, we infer that the combined effects of time domain and frequency domain features on the EEG based biometric recognition are positive.
Table 2. The effect of the feature vector dimension on classification accuracy and speed

| No. | Features | Feature Dimensionality | Classifier | ACC (%) | SPD (sec) |
|-----|----------|------------------------|------------|---------|-----------|
| 1   | Only ApEn | (3)                    | SVM        | 66.60   | 112.120   |
| 2   | Only SampEn | (3)                  | SVM        | 69.90   | 112.250   |
| 3   | Only ShanEn | (3)                  | SVM        | 80.20   | 97.758    |
| 4   | ApEn + SampEn + ShanEn | (6)          | SVM        | 88.10   | 15.658    |
| 5   | Mean and STD [11] | (6)          | SVM        | 84.15   | 78.586    |
| 6   | 1st-4th OS | (12)                  | SVM        | 85.50   | 28.143    |
| 7   | 1st-4th OS + median + mode | (18)       | SVM        | 89.90   | 4.087     |
| 8   | Statistics\(a\) + ApEn | (18)         | SVM        | 93.90   | 1.582     |
| 9   | Statistics\(a\) + SamEn | (18)         | SVM        | 94.50   | 1.611     |
| 10  | Statistics\(a\) + ShanEn | (24)        | SVM        | 95.20   | 1.081     |
| 11  | Statistics\(a\) | (21)             | SVM        | 94.30   | 2.103     |
| 12  | Statistics\(a\) + ApEn,ShanEn | (21)     | SVM        | 94.90   | 1.111     |
| 13  | Statistics\(a\) + SamEn,ShanEn | (21)     | SVM        | 96.00   | 1.089     |
| 14  | CFS       | (23)                  | SVM        | 96.90   | 1.6212    |
| 15  | Statistics\(a\) + ApEn,SamEn,ShanEn | (30)    | SVM        | 95.70   | 4.366     |

\(a\) The Statistics represented the \(M_i^s\), median, mode and z-normalization value.

Furthermore, the higher order statistic features of the distributions can better perform the recognition task than the elementary statistic features, such as mean and standard deviation in our research; and the median, mode, and z-normalization value of the distributions play an important role in overcoming the extreme value issue, and to distinguish the signals’ (dis)similarity. Over 5% classification accuracy rate improved with these statistical features in our proposed method. This table also noted that the subject recognition results could have performed well when extended the statistic features with the chaos features from both time-domain and frequency domain.

Finally, we streamlined the existing feature set by using the CFS algorithm in order to reduce the vector dimension. The third order and fourth order statistical moment on the original data sequence and the second difference of the data sequence are removed because of the high intercorrelation coefficients of those features. In our experiments, the results of the optimized data indicate that the \(d = 23\) dimensional feature vector can achieve the best accuracy performance, while the \(d > 23\), the accuracy performance decrease relatively.

4. Conclusion

In this paper, a feature extraction method based on quantitative measurement has been proposed for EEG recognitions. To investigate the ability of the proposed method, both time-dependent and
frequency-dependent features were evaluated in this study. The CFS algorithm was utilized for the feature selection to find an optimized solution for EEG authentication. The results demonstrated that the proposed method could reduce the heavy computational load and ensures a relatively high classification accuracy when using an SVM classifier to recognize the users under non-resting states. In this study, the classification results estimated a 12.75% enhancement of the AF3 channel, which is more accurate than previous studies using the same dataset. In future research, more work will consider the EEG data collection and the impact of the signal segmentation on practical applications.

References
[1] Paranjape R, Mahovsky J, Benedicenti L and Koles Z 2001 The electroencephalogram as a biometric Electrical and Computer Engineering, 2001. Canadian Conference on vol 2 (IEEE) pp 1363–1366
[2] La Rocca D, Campisi P and Scarano G 2012 Eeg biometrics for individual recognition in resting state with closed eyes Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the (IEEE) pp 1–12
[3] Brigham K and Kumar B V 2010 Subject identification from electroencephalogram (eeg) signals during imagined speech Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on (IEEE) pp 1–8
[4] Abdullah M K, Subari K S, Loong J L C and Ahmad N N 2010 Analysis of effective channel placement for an eeg-based biometric system Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on (IEEE) pp 303–306
[5] Poulos M, Rangoussi M and Alexandris N 1999 Neural network based person identification using eeg features Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on vol 2 (IEEE) pp 1117–1120
[6] Al-Fahoum A S and Al-Fraihat A A 2014 ISRN neuroscience 2014
[7] La Rocca D, Campisi P, Vegso B, Cserti P, Kozmann G, Babiloni F and Fallani F D V 2014 IEEE transactions on Biomedical Engineering 61 2406–2412
[8] Berkhout J and Walter D O 1968 IEEE Transactions on Biomedical Engineering 165–168
[9] Hayes M H 1996 Statistical Digital Signal Processing and Modeling 1st ed (New York, NY, USA: John Wiley & Sons, Inc.) ISBN 0471594318
[10] Picard R W, Vyzas E and Healey J 2001 IEEE Transactions on Pattern Analysis and Machine Intelligence 23 1175–1191
[11] Li S, Cha S H and Tappert C C 2018 Biometric distinctiveness of brain signals based on eeg The 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2018) (IEEE)
[12] Li S, Leider A, Qiu M, Gai K and Liu M 2017 Brain-based computer interfaces in virtual reality International Conference on Cyber Security and Cloud Computing (CSCloud) (IEEE) pp 300–305
[13] Koike-Akino T, Mahajan R, Marks T K, Wang Y, Watanabe S, Tuzel O and Orlik P 2016 Highaccuracy user identification using eeg biometrics 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE) pp 854–858
[14] Srihari S N, Cha S H, Arora H and Lee S 2002 Journal of forensic science 47 1–17
[15] Cha S H and Srihari S 2000 Multiple feature integration for writer verification Proceedings 7th International Workshop on Frontiers in Handwriting Recognition pp 333–342
[16] Klem G H, Lu¨ders H O, Jasper H, Elger C et al. 1999 Electroencephalogr Clin Neurophysiol 52 3–6