# EEG signals similarity based on compression

Michal Prílepok, Jan Platoš, and Václav Snášel

Department of Computer Science, FEECS
IT4 Innovations, European Center of Excellence
VSB-Technical University of Ostrava
Ostrava, Czech Republic
{michal.prilepok, jan.platos, vaclav.snasel}@vsb.cz

**Abstract.** The electrical activity of brain or EEG signal is very complex data system that may be used to many different applications such as device control using mind. It is not easy to understand and detect useful signals in continuous EEG data stream. In this paper, we are describing an application of data compression which is able to recognize important patterns in this data. The proposed algorithm uses Lempel-Ziv complexity for complexity measurement and it is able to successfully detect patterns in EEG signal.

**Keywords:** Electroencephalography; EEG; BCI; EEG waves group; EEG data; LZ Complexity

## 1 Introduction

The Electroencephalography (EEG) plays a big role in diagnosis of brain diseases, and, also, in Brain Computer Interface (BCI) system applications that helps disabled people to use their mind to control external devices. Both research areas are growing today.

The EEG records the electrical activity of the brain using several sensors placed on a scalp . Different mental tasks produce indiscernible recordings but they are different. Different brain actions activate different parts of the brain. The most difficult part is the definition of an efficient method or algorithm for detection of the differences in recordings belonging to the different mental tasks. When we define such algorithm we are able to translate these signals into control commands of an external device, e.g. prosthesis, wheelchair, computer terminal, etc.

## 2 The Electroencephalography

The Electroencephalography (EEG) measures the electrical activity of human brain, by placing set of sensors on a scalp, according to 10/20 EEG International electrode placement, as is depicted on Figure 1. The measuring of EEG signal records can be done between two active electrodes (bipolar recording), or between an active electrode and a reference electrode (mono-polar recording) [16].

J. Pokorný, K. Richta, V. Snášel (Eds.): Dateso 2014, pp. 59–70, ISBN 978-80-01-05482-6.

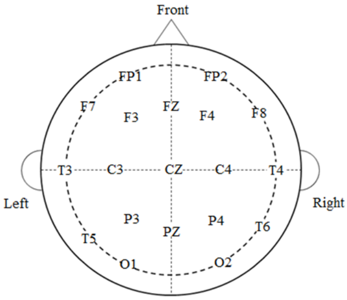

**Fig. 1.** 10/20 International Electrode Placement

### 2.1 EEG Waves Types

The types of brain waves distinguished by their different frequency ranges are recognized as follows.

– Delta ($\delta$) waves lie within the range from approximately 0.5 up to 4 Hz. The amplitude of this waves is varying and have been associated with deep sleep and present in the waking state.
– Theta ($\theta$) waves lie within the range from 4 to 7.5 Hz. The amplitude varies about 20 $\mu$V. Theta waves have been associated with access to unconscious material, creative inspiration and deep meditation.
– The frequency of the Alpha ($\alpha$) waves lies within the range from 8 to 13 Hz, the amplitude varies between 30 and 50 $\mu$V. It is reduced or eliminated by opening the eyes, by hearing unfamiliar sounds, by anxiety, or mental concentration or attention.
– Beta ($\beta$) waves are the electrical activities of the brain varying within the frequency range from 14 to 26 Hz. The amplitude is about 5 up to 30 $\mu$V. Beta waves has been associated with active thinking, active attention, focus on the outside world, or solving concrete problems. A high-level beta wave may be acquired when a human is in a panic state.
– Gamma ($\gamma$) waves have frequency range above 30 Hz, can be used to demonstrate the locus for right and left index finger movement, right toes, and the rather broad and bilateral area for tongue movement [19, 18].
– Mu ($\mu$) waves will be same Alpha frequency range 8 to 13 Hz, but Alpha waves are recorded on occipital cortex area, and Mu waves are recorded on motor cortex area. Mu waves are related to spontaneous nature of the brain such motor activities [18].

### 2.2 History of EEG

Carlo Matteucci and Emil Du Bois-Reymond, were first people who register the electrical signals emitted from muscle nerves using a galvanometer and established the concept of neurophysiology. The first brain activity in the form of electrical signals was recorded in 1875, by Richard Caton (1842–1926), a scientist from Liverpool, England, using a galvanometer and two electrodes placed over the scalp of a human. From here EEG stand to, Electro that referring to registration of brain electrical activities, Encephalon that referring to emitting the signals from a brain, and gram or graphy, which means drawing. Then the term EEG was henceforth used to denote electrical neural activity of the brain [19].

In 1920, Hans Berger, the discoverer of the existence of human EEG signals, began his study of human EEG. In 1910, Berger started working with a string galvanometer and later he used a smaller Edelmann model. After the year 1924, he used larger Edelmann model. Berger started to use the more powerful Siemens double coil galvanometer (attaining a sensitivity of 130 $\mu$ V/cm) in 1926. In 1929 Berger made the first report of human EEG recordings with duration from one to three minutes on photographic paper and, in the same year, he also found some correlation between mental activities and the changes in the EEG signals [19].

The first biological amplifier for the recording of brain potentials was built by Toennies (1902–1970). In 1932 the differential amplifier for EEG recording was later produced by the Rockefeller foundation. The potential of a multichannel recordings and a large number of electrodes to cover a wider brain region was recognized by Kornmuller. Berger assisted by Dietch (1932) applied Fourier analysis to EEG sequences, which was developed during the 1950s [19].

After that the EEG analysis and classification take grow and development every day. The application of the EEG signals to diagnosis of the brain diseases and to control external devices for disabled people such as wheel chair, prosthesis, etc. Today, several techniques for analysis and classification the EEG signal exists, by using EEG multichannel recording according to 10/20 International electrodes standard, which is used in Brain Computer Interface (BCI).

## 3 Related works

In this section we present some of related works for EEG data analysis using different techniques such as Non-negative Matrix Factorization (NMF), Normalized Compression Distance (NCD), and Lempel-Ziv (LZ) complexity measure, and Curve Fitting (CF).

Lee et al. presented a Semi-supervised version of NMF (SSNMF) which jointly exploited both (partial) labeled and unlabeled data to extract more discriminative features than the standard NMF. Their experiments on EEG datasets in BCI competition confirm that SSNMF improves clustering as well as classification performance, compared to the standard NMF [10].

Shin et al. have proposed new generative model of a group EEG analysis, based on appropriate kernel assumptions on EEG data. Their proposed model finds common patterns for a specific task class across all subjects as well as individual patterns that capture intra-subject variability. The validity of the proposed method have been tested on the BCI competition EEG dataset [20].

Dohnalek et al. have proposed method for signal pattern matching based on NMF, also they used short-time Fourier transform to preprocess EEG data and

Cosine Similarity Measure to perform query-based classification. This method of creating a BCI capable of real-time pattern recognition in brainwaves using a low cost hardware, with very cost efficient way of solving the problem [5]. In this context, Gajdos et al. implemented the well-performing Common Tensor Discriminant Analysis method [6] using massive parallelism [7].

Mehmood, and Damarla applied kernel Non-negative Matrix Factorization to separate between the human and horse footsteps, and compared KNMF with standard NMF, their result conclude that KNMF work better than standard NMF [14].

Sousa Silva, et al. verified that the Lempel and Ziv complexity measurement of EEG signals using wavelets transforms is independent on the electrode position and dependent on the cognitive tasks and brain activity. Their results show that the complexity measurement is dependent on the changes of the pattern of brain dynamics and not dependent on electrode position [4].

Noshadi et al. have applied Empirical mode decomposition (EMD) and improved Lempel-Ziv (LZ) complexity measure for discrimination of mental tasks, their results reached 92.46% in precision, and also they concluded that EMD-LZ is getting better performance for mental tasks classification than some of other techniques [15].

Li Ling, and Wang Ruiping calculated complexity of sleeping stages of EEG signals, using Lempel-Ziv complexity. Their results showed that nonlinear feature can reflect sleeping stage adequately, and it is useful in automatic recognition of sleep stages [13].

Krishna, et al. proposed an algorithm for classification of the wrist movement in four directions from Magnetoencephalography (MEG) signals. The proposed method includes signal smoothing, design of a class-specific Unique Identifier Signal (UIS) and curve fitting to identify the direction in a given test signal. The method was tested on data set of the BCI competition, and the best result of the prediction accuracy reached to 88.84 % [9].

Klawonn, et al. have applied Curve Fitting for Short Time Series biological data to remove noise from measured data and correct measurement errors or deviations caused by biological variation in terms of a time shift etc. [8]

## 4 Similarity

The main property in the similarity is a measurement of the distance between two objects. The ideal situation is when this distance is a metric [21]. The distance is formally defined as a function over Cartesian product over set $S$ with non-negative real value (see [3, 12]). The metric is a distance which satisfy three conditions for all:

**Definition 1.** *A mapping $D : U \to \mathbb{R}^+$ is said to be a distance on the universe $U$ if the following properties hold:*

*D1 Non-negativity: $D(x, y) \geq 0$ for any $x, y \in U$;*
*D2 Symmetry: $D(x, y) = D(y, x)$ for any $x, y \in U$;*

*D3 Identity of indiscernibles:* $D(x, y) = 0$ *if and only if* $x = y$*;*
*D4 Triangular inequality:* $D(x, y) \leq D(x, z) + D(z, y)$ *for any* $x, y, z \in U$*.*

### 4.1 Lempel-Ziv Complexity

The Lempel-Ziv (LZ) complexity for sequences of finite length was suggested by Lempel and Ziv [11]. It is a non-parametric, simple-to-calculate measure of complexity in a one-dimensional data. LZ complexity is related to the number of distinct substrings and the rate of their recurrence along the given sequence [17], with larger values corresponding to more complexity in the data. It has been applied to study the brain function, detect ventricular tachycardia, fibrillation and EEG [22]. It has been applied to extract complexity from mutual information time series of EEGs in order to predict response during isoflurane anesthesia with artificial neural networks [2]. LZ complexity analysis is based on a coarse-graining of the measurements, so before calculating the complexity measure $c(n)$, the signal must be transformed into a finite symbol sequence. In this study, we have used turtle graphic for conversion of measured data into finite symbol sequence $P$. The sequence $P$ is scanned from left to right and the complexity counter $c(n)$ is increased by one unit every time a new subsequence of consecutive characters is encountered. The complexity measure can be estimated using the algorithm described in [11, 2].

In our experiment we do not deal with the measure of the complexity. We create a list of the LZ sequences from the individual subsequence. One list is created for each data file with turtle commands of the compared files.

The comparison of the LZ sequence lists is the main task. The lists are compared to each other. The main property for comparison is the number of common sequences in the lists. This number is represented by the $s_c$ parameter in the following formula, which is a metric of similarity between two turtle commands lists.

$$SM = \frac{s_c}{min(c_1, c_2)} \tag{1}$$

Where

- $s_c$ – count of common string sequences in both dictionaries.
- $c_1, c_2$ – count of string sequences in dictionary of the first or the second data trial.

The $SM$ value is in the interval between 0 and 1. The two documents are equal if $SM = 1$ and they have the highest difference when the result value of $SM = 0$.

## 5 Dataset

The data for our experiments was recorded in our laboratory. We have used 7 channels from recorded data. The signal data contains records of the movement of one finger from four different subjects. Every subject performed a press of a button with left index finger. The sampling rate was set to 256 Hz. The signals were band pass filtered from 0.5 Hz to 60 Hz to remove unwanted lower and higher frequencies and noise. The data was then processed, that we extract each movement from the data as well as 0.3s before the movement and 0.3s after the movement.

The pre-processed data contains 4606 data trials – 2303 data trails with finger movement and 2303 trails without finger movement. We divided it into seven groups, one group for each sensor. In our experiment we are using 75% of data for training and 25% for testing. Each group contains part of training and testing data part. The training part for one sensor contains 492 trials – 246 data trails with finger movement and 246 trails without finger movement. The testing part contains 166 trails – 83 trails with finger movement and 83 trails without finger movement. The we have used for further model validation.

### 5.1 Interpolation of the EEG data

After recording and filtering of the EEG data we apply polynomial curve fitting for data smoothing. The fitting will remove noise from the data and fit the data trend.

Consider the general form for a polynomial fitting curve of order $j$:

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_j x^j = \sum_{k=1}^{j} a_k x^k \tag{2}$$

We minimized the total error of polynomial fitting curve with least square approach. The general expression for any error using the least squares approach is:

$$err = \sum (d_j)^2 \tag{3}$$

$$err = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots + (y_j - f(x_j))^2 \tag{4}$$

$$err = \sum_{i=1}^{n} \left( y_i - \left( a_0 + \sum_{k=1}^{j} a_k x^k \right) \right)^2 \tag{5}$$

where:

- $n$ is count of data points in one move,
- $i$ is the current data point being summed,
- $j$ is the polynomial order.

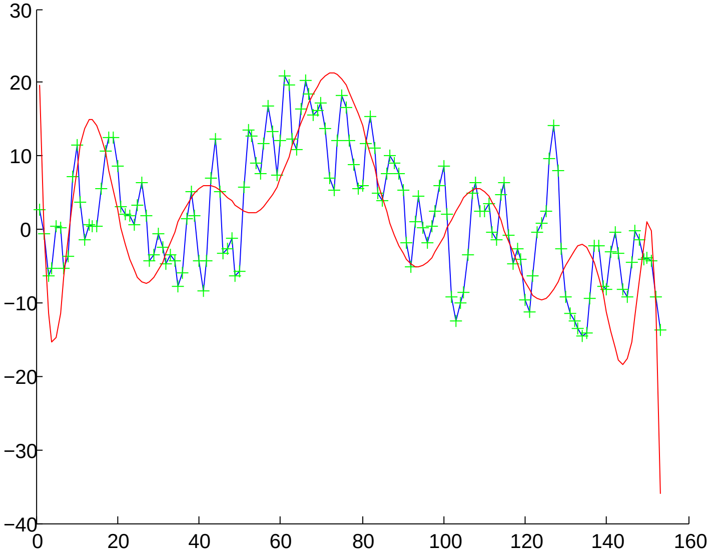

**Fig. 2.** EGG trail before (blue line) and after smoothing (red line) with 15th order polynomial curve fitting.

### 5.2 Turtle Graphics

Consider we have control on a turtle on computer screen, this turtle must be respond on a sequence of commands. These commands: forward command, is moving the turtle in front direction a few number of units, right command rotate turtle in clockwise direction a few number of degrees, Back command and Left command are cause same movement but in opposite way. The number of commands to determine, how much to move is called input commands, depending on the application. When moving the turtle under input commands it leave trace, this trace represent the desired object, as in Figure 3. represent simple example for drawing on screen by steering the turtle with four commands forward, right, left, and back command [1]. By this way can represent and drawing the objects, from simple to complex objects.

## 6 EEG Experiment

The recorded data trail were filtered with band pass filter and divided into individual sensor trails. For each trial we calculated polynomial fitting curve with 15th order and total error minimization with least square approach. The 15th order is enough flexible to smooth data, remove unwanted noise, and to keep the trend of data. After smoothing data we converted calculated curve values into text using Turtle graphics. For the turtle we used 128 commands in two right quadrants – first and fourth. Each command represents one angle – a data

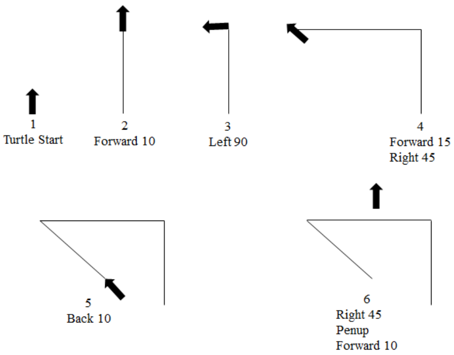

**Fig. 3.** Simple sequence of Turtle Commands

trend direction. We used only two quadrants – the first and fourth, because the time line goes from left to right and the signal does not go backwards into past.

### 6.1 Lempel-Ziv Complexity

After this steps, were prepared a LZ subsequences list from turtle graphics commands list from previous step using LZ complexity for each test EEG trial. Similarities to all train trails using Eq. 1 were calculated for every test trail. Then we selected a group of training trails with similarity $S$ satisfying following condition $S \geq T_{min} \wedge S \leq T_{max}$ for every test trail. The condition threshold values are depicted in Table 1 for all sensors. This selected group of trials is used for calculation in which category belongs the tested trial. This was calculated as a ratio of trials with movement to total count of selected trials in group, using the formula:

$$C = \frac{m_t}{c_t}$$

Where:

- $m_t$ is a count of trails, which are marked as trail with movement,
- $c_t$ is a count of trials in selected group, which satisfy condition.

The tested trail is marked as trail, which belongs to category with movement trails if $C \geq 0.5$ and as a trail without movement otherwise. These steps were performed separately for all categories of data – with movement and without movement – and all sensors.

The values of $T_{min}$ and $T_{max}$ represent the shortest range $R$ in which classifier has correctly identified maximum trials of both categories, with movement and without movement, with emphasis to maximum correctly identified trial with movement, where $T_{min} \in [0, 1]$ and $T_{max} \in [0, 1]$ and $T_{min} < T_{max}$, for example:

$$R(T_{min}, T_{max}) \in [0.15, 0.2]$$

Figure 4 shows a distribution of individual similarities for a trail with (blue bars) and without movement (orange bars). We can see that each data category can be divided into one group. This two groups have with a small intersection between $T_{min}$ and $T_{max}$ value.

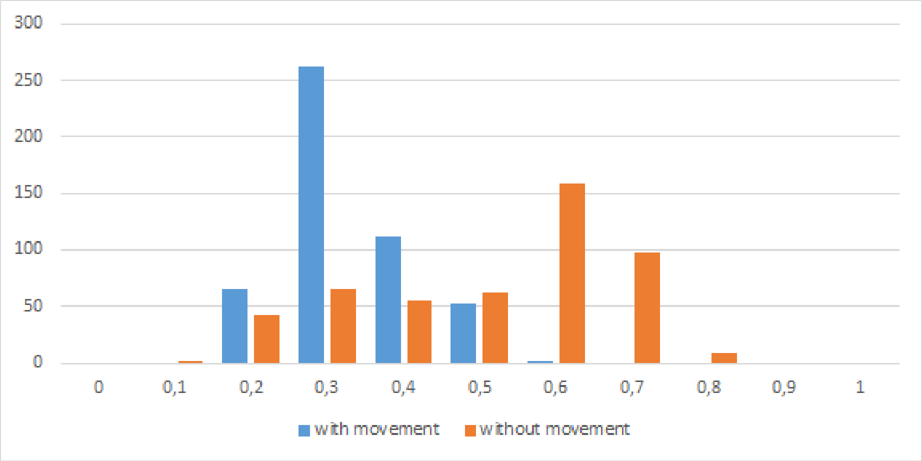

**Fig. 4.** Histogram of the similarities for trial with and without movement of one sensor

### 6.2 Experiment Result

Our experiment is focused on successful detection of both data categories, data with movement and data trial without movement. Our data was divided into seven data parts. Each part contains trails from one sensor. Each data parts has two subparts. The first data subpart contains training data – 75% of trials with movement and without movement. The other part is used as testing data sub part. This is used for our model validation. It contains 25% of trials with movement and without movement.

In our experiment we are able to detect movement of index finger with success detection rate between 56.02% and 58.78%. The best results we reach up on sensor S5 (58.78%) and S2, S4 (58.43%). The worst result is for sensor S7 (56.02%). The detection results and their corresponding threshold values for all sensor are in Table 1.

Detection rate in trials with movement varies between 36.14% (S6) and 72.28% (S7). Detection rate in trials with no movement varies between 39.75% (S7) and 77.10% (S6).

Most of the values taken by *minThreshold* are around 0.30 and *maxThreshold* values are situated around value 0.50.

**Table 1.** Table of Results

| **Sensor** | $T_{min}$ | $T_{max}$ | **Detection rate** | | **Total detection rate** |
|---|---|---|---|---|---|
| | | | **Movement** | **No movement** | |
| **S1** | 0.40 | 0.65 | 61.44% | 51.80% | 56.62% |
| **S2** | 0.35 | 0.45 | 60.24% | 56.62% | 58.43% |
| **S3** | 0.30 | 0.45 | 59.03% | 55.42% | 57.22% |
| **S4** | 0.30 | 0.60 | 66.26% | 50.60% | 58.43% |
| **S5** | 0.30 | 0.40 | 49.39% | 68.29% | 58.78% |
| **S6** | 0.60 | 0.65 | 36.14% | 77.10% | 56.62% |
| **S7** | 0.25 | 0.50 | 72.28% | 39.75% | 56.02% |

## 7 Conclusion

We made our experiments on our EEG data recorded in our laboratory from four different subjects performing the same task – pressing a button with index finger. The EEG data was recorded using 7 channels recording machine with sampling frequency 256 Hz. The signals were band pass filtered from 0.5 Hz to 60 Hz to remove unwanted frequencies and noise. The signals record the movement of one finger. After removing unwanted frequencies and noise we preprocessed data with polynomial curve fitting with 15th order, turtle graphic – conversion from number into text and Lempel-Ziv complexity – similarity measurement.

In this paper we applied a successful approach for index finger movement detection. Our suggested approach use polynomial fitting curve for smoothing recorded data and Lempel-Ziv complexity for measuring similarity between trails. Our approach is able to correctly detect EEG trail of index finger with success rate between 56.02% and 58.78%. The best results we reach up on sensor 58.78% and 58.43%. The worst result is for sensor 56.02%. Detection rate in trials with movement varies between 36.14% and 72.28%. Detection rate in trials with no movement varies between 39.75% and 77.10% .

The method proposed in this work seems to be able to detect trails with and without movement with overall successful rate more than 56.02%. It can be applied to the use on real data.

## Acknowledgment

This work was partially supported by the Grant of SGS No. SP2014/110, VŠB-Technical University of Ostrava, Czech Republic, and was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

## References

1. H. Abelson and A. diSessa. *Turtle Geometry: The Computer as a Medium for Exploring Mathematics.* The MIT Press, July 1986.
2. D. Abásolo, R. Hornero, C. Gómez, M. García, and M. López. Analysis of EEG background activity in alzheimer's disease patients with lempel–ziv complexity and central tendency measure. *Medical Engineering & Physics*, 28(4):315 – 322, 2006.
3. R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.
4. A. de Sousa Silva, A. Arce, A. Tech, and E. Costa. Quantifying electrode position effects in eeg data with lempel-ziv complexity. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 4002–4005, 2010.
5. P. Dohnálek, P. Gajdoš, T. Peterek, and M. Penhaker. Pattern recognition in EEG cognitive signals accelerated by GPU. volume 189 AISC, pages 477–485. 2013. cited By (since 1996)1.
6. A. Frolov, D. Husek, and P. Bobrov. Brain-computer interface: Common tensor discriminant analysis classifier evaluation. In *Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on*, pages 614–620, 2011.
7. P. Gajdos, P. Dohnalek, and P. Bobrov. Common tensor discriminant analysis for human brainwave recognition accelerated by massive parallelism. In *Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on*, pages 189–193, 2013.
8. F. Klawonn, N. Abidi, E. Berger, and L. Jänsch. Curve fitting for short time series data from high throughput experiments with correction for biological variation. In J. Hollmén, F. Klawonn, and A. Tucker, editors, *IDA*, volume 7619 of *Lecture Notes in Computer Science*, pages 150–160. Springer, 2012.
9. S. Krishna, K. Vinay, and K. B. Raja. Efficient meg signal decoding of direction in wrist movement using curve fitting (emdc). In *Image Information Processing (ICIIP), 2011 International Conference on*, pages 1–6, 2011.
10. H. Lee, J. Yoo, and S. Choi. Semi-supervised nonnegative matrix factorization. *Signal Processing Letters, IEEE*, 17(1):4–7, 2010.
11. A. Lempel and J. Ziv. On the complexity of finite sequences. *Information Theory, IEEE Transactions on*, 22(1):75–81, 1976.
12. M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.
13. L. Ling and W. Ruiping. Complexity analysis of sleep eeg signal. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pages 1–3, 2010.
14. A. Mehmood and T. Damarla. Kernel non-negative matrix factorization for seismic signature separation. *Journal of Pattern Recognition Research*, 8(1):13–25, 2013.
15. S. Noshadi, V. Abootalebi, and M. Sadeghi. A new method based on emd and lz complexity algorithms for discrimination of mental tasks. In *Biomedical Engineering (ICBME), 2011 18th Iranian Conference of*, pages 115–118, 2011.
16. R. Q. Quiroga. *Quantitative analysis of EEG signals: Time-Frequency methods and Chaos Theory.* PhD thesis, Institute of Signal Processing and Institute of Physiology, Medical University of Lubeck, Germany, 1998.
17. N. Radhakrishnan and B. Gangadhar. Estimating regularity in epileptic seizure time-series data. *Engineering in Medicine and Biology Magazine, IEEE*, 17(3):89–94, 1998.

18. T. K. Rao, M. R. Lakshmi, and T. V. Prasad. An exploration on brain computer interface and its recent trends. *International Journal of Advanced Research in Artificial Intelligence*, 1(8):17 – 22, 2012.
19. S. Sanei and J. Chambers. *EEG Signal Processing*. John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2007.
20. B. Shin and A. Oh. Bayesian group nonnegative matrix factorization for eeg analysis. *CoRR*, abs/1212.4347:1–8, 2012.
21. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. cited By (since 1996)1968.
22. X.-S. Zhang, R. Roy, and E. Jensen. Eeg complexity as a measure of depth of anesthesia for patients. *IEEE Transactions on Biomedical Engineering*, 48(12):1424–1433, 2001. cited By (since 1996)165.