Published in final edited form as:

Proc Int Jt Conf Neural Netw. 2013 ; : 1–8. doi:10.1109/IJCNN.2013.6707106.

A Framework for Content-based Retrieval of EEG with Applications to Neuroscience and Beyond*

Kyungmin Su and

Computer Science Department, University of Texas at San Antonio, San Antonio, TX 78249 USA
(phone: 210-458-7662; fax: 210-458-4437)

Kay A. Robbins

Computer Science Department, University of Texas at San Antonio, San Antonio, TX 78249 USA

Kyungmin Su: ksu@cs.utsa.edu; Kay A. Robbins: Kay.Robbins@utsa.edu

Abstract

This paper introduces a prototype framework for content-based EEG retrieval (CBER). Like content-based image retrieval, the proposed framework retrieves EEG segments similar to the query EEG segment in a large database.

Such retrieval of EEG can be used to assist data mining of brain signals by allowing researchers to understand the association between brain patterns, responses, and the environment. Retrieval might also be used to enhance the accuracy of Brain Computer Interface (BCI) systems by providing related samples for training.

We present key components of CBER and explain how to handle the distinctive characteristics of EEG. To demonstrate the feasibility of the framework, we implemented a simple EEG database of about 37,000 samples from more than 100 subjects. We ran two retrieval scenarios with a set of EEG features and evaluation metrics. The results of finding similar subjects clearly demonstrate the potential of CBER in many EEG applications.

I. Introduction

Content-based retrieval, sometimes known as query-by content, incorporates methods of computer vision and machine learning to formulate queries based on the content of a data item rather than its labeling or metadata. Content-based image retrieval (CBIR) has had a long and successful history [1], with applications in many areas including automatic identification from surveillance video, automatic tagging or annotation, and enhanced search of online image databases.

*This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

In medicine, content-based retrieval has important applications to CAD (computer aided medical diagnosis). For example, Alto et al. [2] report on the development of content-based image retrieval of mammograms to aid in the diagnosis of breast cancer. The metadata associated with the retrieved items can provide a context of potential outcomes to aid radiologist decision-making and recommendations for treatment. More sophisticated approaches using context and metadata as well as image features have been shown to improve diagnosis [3].

Content-based retrieval has also been applied to other types of signals. For example, Pauer et al. [4] created a database of EMG sensor data to retrieve stored segments in similar body poses, and Harris et al. [5] implemented content-based audio retrieval for speaker recognition.

In bioinformatics, hundreds of databases that support content-based retrieval of gene expression and/or other information are widely available and linked in an extensive network of interrelated information. The value of these systems in advancing understanding of the genome is indisputable [6], [7].

In this paper, we propose a framework for content-based Electroencephalography (EEG) retrieval (CBER). The idea is that given a small segment of EEG, one might perform a query to retrieve similar segments in a larger corpus. Such retrieval in a large database of EEG can provide a context to aid in analysis. For example, a researcher can examine the metadata for the set of responses to a query to determine what the responses have in common. If certain metadata values occur in the response more frequently than random (i.e., the response is enriched for a particular value), one might potentially make an association of the query with the metadata value. Such enrichment approaches are widely used in bioinformatics to annotate unknown genes and other entities.

Content-based EEG retrieval (CBER) might also be used to enhance the accuracy of Brain Computer Interface (BCI) systems. Lu et al. [8] proposed the regularized CSP to overcome the small sample problem of BCI classifier training. They use samples from other randomly selected subjects to increase the number of training samples. A database supporting similar-subject retrieval could provide samples of similar subjects to enhance the classifier performance rather than randomly selected subjects.

Large databases of EEG are also becoming available for computer-aided diagnosis of epilepsy. The EPILEPSIAE project is a European effort to produce a high-quality, highly-annotated database of multi-channel continuous EEG recordings of 300 subjects that includes epileptic events [9]. A CBER front end could look for similar epileptic events, but can also retrieve patients with similar normal brain signals.

Finally, CBER approaches can be used to assist data mining of brain signals by allowing researchers to understand the association between brain patterns, responses, and the environment.

CBER has only recently become feasible and presents a number of technical challenges beyond those faced by more traditional image-based systems. Foremost is the lack of

available ground truth. An ordinary person can look at an image and evaluate whether it resembles a Starbuck's logo, but identification of EEG signals beyond a few stereotypical behaviors associated with artifacts or brain states such as alpha spindling, is not possible from visual inspection. Another issue is the lack of availability of large-scale annotated databases. A third issue is the difficulty of framing questions that can be answered to further understanding of signals.

A content-based EEG retrieval system, like its CBIR counterpart, starts with a corpus of data, a set of feature functions, a retrieval mechanism, and an evaluation process. A simple example would be to use the raw signal as a feature vector and simply retrieve vectors that are closest to this vector. Either a distance cutoff or the number of results returned could limit the search. Such an approach could be easily supported by existing database infrastructure such as PostgreSQL [10].

A difficulty with such a straightforward approach is that EEG signals are noisy and subject to artifacts. There is no standardized way of removing artifacts from either the stored signals or the query signals. The choice of removal approach might have a significant impact on the signals that are retrieved. Also, the signal amplitudes depend on the exact placement of the headsets making straightforward distance metrics problematic, and high-dimensional nature of the query features will suffer the curse of dimensionality.

This paper introduces a framework for CBER and demonstrates its feasibility. We use a publicly available EEG corpus (the BCI2000 dataset available on PhysioNet [11], [12]), some simple, frequently-used features, and several established metrics to retrieve EEG segments belonging to the same and similar subjects. Section II introduces the features and the metrics. Section III describes the experimental setup, and Section IV presents the experimental results. Section V discusses these results and the prospects for CBER.

II. CBER Framework

The basic workflow for CBER is shown in Fig. 1. Preprocessing and feature extraction must be accessible during both database construction and query processing. Preprocessing of EEG data generally includes filtering, faulty signal removal, and artifact removal. In this paper, we have chosen features that can easily be computed on-the-fly on short segments of data. This requirement is not strictly necessary for data mining applications, where large segments of input EEG might be mined by comparison with a database. Depending on the application, the retrieved data might be post processed, possibly by combining with information from other sources or by statistically refining retrieved results. The remainder of this section introduces the features and the metrics used for evaluation in this paper.

A. Features

For feature extraction, we assume that each EEG dataset contains measurements from K segments or epochs, containing N time points measured from each of L channels. The data is presented as an $L \times N \times K$ array of epoched data. A query consists of an $L \times N$ array containing the measurements from a single epoch. In this paper, we consider five (5) frequently used EEG feature types as listed in Table I.

The features shown in Table I fall into three categories: statistical, temporal, and spectral. All of these features are amplitude-independent, in the sense that multiplying the signal by a constant after removing the mean does not change the value. The resulting feature sizes, when calculated from an $L \times N$ input vector, are listed in the last column of Table I. Table II summarizes some notation needed for describing the features.

Kurtosis, also known as the scaled 4th moment, measures the size of the tails in the signal probability distribution. For the time series X , the kurtosis is defined as:

$$f_K(X) = \frac{M_4}{\sigma_x^2}$$

The kurtosis function maps a vector into a single value. When applied to an array of $L \times N$ inputs, the kurtosis is evaluated for each row, i.e. channel, separately to calculate a feature vector of size L . The kurtosis of the normal distribution is 3. Kurtosis values greater than 3 indicate the probability distribution of X has tails that contain more weight than a normal distribution. Kurtosis is useful for detecting a preponderance of uncommonly large or small values and is often used in EEG data for artifact detection, particularly of eye blinks, which have large amplitude deviations from normal signals.

The mobility, which is defined as the ratio of the standard deviation of the first difference and standard deviation of the original signal, is a measure of the relative average slope. For the time series X , the mobility is defined as [13], [14]:

$$f_M(X) = \frac{\sigma_D}{\sigma_X}$$

Like the kurtosis, the mobility function maps a vector into a single value and generates a feature vector of size L for an $L \times N$ input array. In [13], Hjorth interprets the mobility in the frequency domain and shows that the mobility is directly related to the standard deviation of the power spectrum along the frequency axis. Therefore, the mobility can be used to compare the spectral characteristics of EEG data.

The complexity, which is the ratio of the mobility of the first difference of the signal and the mobility of the signal itself, measures the amount of detail in the time series signal. The mobility of the first difference is obtained in a way analogous to the mobility of the original signal, but uses the first and second differences of the signals, respectively. For the time series X , the complexity is defined as [13], [14]:

$$f_C(X) = \frac{f_M(D)}{f_M(X)} = \frac{\sigma_S \cdot \sigma_X}{\sigma_D^2}$$

For an $L \times N$ input array, the complexity generates a feature vector of size L . The minimum value of the complexity is unity when the signal does not have any details. An example of a signal with no details is a pure sine wave [13]. Adding details increases the complexity from

unity. However, the complexity is not a linear function of the number of sinusoidal components. Although mobility is a nonlinear function, it is still useful for comparison of EEG signals.

The autoregressive (*AR*) method represents the time signal as a linear combination of the signal values at the previous P time points:

$$x_n = \sum_{i=1}^P \alpha_i x_{n-1} + \varepsilon_n$$

where ε_n represents noise and the α_i are computed by using linear regression to fit the data. For the time series X , the *AR* feature vector using model order P is defined as:

$$f_A(X) = [\alpha_1 \quad \cdots \quad \alpha_P]'$$

Because the *AR* model maps a vector into P values, it generates a feature vector of size $L \times P$ for an input array of size $L \times N$. We use the *AR* function to extract dynamic information for each EEG channel.

The band intensity ratio (B_{IJ}) is a ratio of spectral intensity in frequency band I to that in frequency band J . Following [15] we use the sum of absolute values of the Fourier coefficients in each frequency band to represent the information. In EEG research, δ (0.5–4 Hz), θ (4–7 Hz), α (8–12 Hz), β (12–30 Hz), and γ (30–100 Hz) are commonly used frequency ranges. In the paper we use $f_B = B_{\alpha\beta}$, the ratio of intensities in the alpha (α) and beta (β) bands because these bands are known to be related to decision-making and motor activity.

B. Evaluation metrics

To measure the performance of retrieval, we compare three evaluation metrics: Mean Average Precision (*MAP*), Averaged Normalized Modified Retrieval Rank (*ANMRR*), and Mean Normalized Retrieval Order (*MNRO*).

MAP is a set-based evaluation metric, while *ANMRR* and *MNRO* are rank-based evaluation metrics. Set-based metrics are useful when a user only cares whether the correct samples are retrieved but is not concerned about retrieval order [16]. For example, in image retrieval, the order of the retrieved samples may not be important because a user intends to evaluate the entire result set. On the other hand, if results retrieved first are weighted more heavily or treated as more relevant, then a ranked-based metric should be used. *ANMRR* is a rank-based metric that was adopted in the MPEG-7 standard [17]. *MNRO* is another rank-based metric that accounts for the database size and relative frequencies of the image content. Depending on the application type, users can choose an appropriate metric for their evaluation.

The precision is the ratio of the number of correct or relevant items retrieved over the total number retrieved. The average precision *AP* for a single query q is the average of the precision scores after each correct retrieved item [18].

$$AP(q) = \frac{1}{N_R(q)} \sum_{i=1}^{N_R(q)} P_q(r_i)$$

where $P_q(r_i)$ is the precision at the i^{th} retrieved relevant item, r_i , and $N_R(q)$ is the number of relevant samples for query q . MAP is the mean of the average precision scores over all queries in the query set Q :

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)$$

Rank-based measures are often used to evaluate retrieval because this metric assumes that correct queries retrieved earlier (ranked first) are of more value [16]. $ANMRR$, a retrieval metric used by MPEG-7, is a rank-based metric that normalizes its score into the range of 0 to 1 for the best to the worst retrieval results, respectively [17].

Simple rank-based measures can be affected by outlier relevant samples. To reduce the effect of outliers, $ANMRR$ defines a threshold K , the boundary of meaningful results, as:

$$K = \min(X \times N_R(q), 2 \times G)$$

where $G = \max\{N_R(q)\}$ and X is 2 if $N_R(q) > 50$, otherwise X is 4. $AVR(q)$, the average retrieval rank score of query q , is defined as:

$$AVR(q) = \frac{1}{N_R(q)} \sum_{i=1}^{N_R(q)} s(r_i)$$

where $s(r_i)$ is the score of i^{th} relevant sample r_i . If the rank of r_i is smaller than K , the score is the rank of r_i . Otherwise, the score is $1.25 \times K$. In other words, the AVR applies a rank cutoff to reduce the effect of outliers. The modified retrieval rank score (MRR) and the normalized modified retrieval rank ($NMRR$) are defined as:

$$MRR(q) = AVR(q) - 0.5 \times [1 + N_R(q)]$$

$$NMRR(q) = \frac{MRR(q)}{1.25 \times K - 0.5 \times [1 + N_R(q)]}$$

Finally, $ANMRR$ is the average of $NMRR$ over all queries. Due to the threshold K , $ANMRR$ is less affected by outlier relevant samples. However, the $ANMRR$ does not account for the size of the database.

Chatzichristofis et al. [16] proposed a modification of $ANMRR$ called $MNRO$ that adjusts the K factor to account for the size of the database and then applies a sigmoidal scaling to calculate a nonlinear rank weighting that rises quickly for ranks less than K and then tapers off.

To count for the size of database in the evaluation, *MNRO* defines K according to the generality of the query. Let the query generality, $g(q)$, denote the generality of the query, which is the fraction of relevant samples in the database.

$$g(q) = \frac{N_R(q)}{N}$$

Here $N_R(q)$ is the total number of relevant samples to the query q , and N is the total number of samples in a database. A large value of generality means that it is easy to find relevant samples in a database because there are many relevant samples. Small generality means there are only few relevant samples or the database is very large, so it is harder to find relevant samples.

The adjusted value of K considers both the number of relevant samples and the size of the database:

$$K = \begin{cases} 4 \times N_R(q) & g(q) \geq 0.01 \\ \frac{0.04}{g(q)} \times N_R(q) & g(q) < 0.01 \end{cases}$$

The Normalized Retrieval Order (*NRO*) for the correct item r is defined as:

$$NRO(r) = \begin{cases} 0 & GT(r) = Rank(r) \\ e^{-9.3668 \times e^{-5.2074 \times ARank(r)}} & GT(r) < Rank(r) \end{cases}$$

Here $GT(r)$ is the rank of retrieved sample r in the correct samples, and $ARank(r)$ is:

$$ARank(r) = \frac{Rank(r) - 1}{K - 1}$$

$MNRO(q)$ is the mean over the set of relevant samples $R(q)$ for query q :

$$MNRO(q) = \frac{1}{N_R(q)} \sum_{r \in R(q)} NRO(r)$$

$MNRO$ is the mean of $MNRO(q)$ over all queries.

III. Experimental Setup

Fig. 1 gives an overview of the workflow for our EEG retrieval system, and Table III summarizes the parameter choices used in this paper to illustrate the framework. The details of each step are explained in the following sections.

A. Experimental data

We use the BCI2000 dataset collection (EEG Motor Movement/Imagery Dataset) created by Schalk et al. [11]. The collection is publicly available and can be downloaded from PhysioNet [12]. This data collection consists of 64-channel EEG recordings from 109 subjects. Each subject performed two baseline tasks (eyes open and eyes closed) and four different types of motor imagery task pairs (open or close left/right fist, imagine open or close left/right fist, open or close fists/hands, and imagine open or close fists/hands). A baseline task dataset contains one long (1 minute) baseline epoch and a motor imagery task dataset contains thirty short (4.1 ~ 4.2 seconds) epochs that are a specific task pair intermixed with a rest task. Because each subject did one experimental run of each baseline task one time and three runs of each of the four pairs of motor imagery tasks, each subject has a total of 14 datasets and a total of 362 epochs.

We excluded subjects 88, 92, and 100 because they were acquired at a different sampling rate and subject 89 because of non-standard annotations. We also excluded subjects 104 and 106 because they have some shorter epochs than other subjects. The test database therefore consisted of 37,286 epochs from 103 subjects.

B. Removal of noise and artifacts

To remove the overall signal mean and general signal noise, we use several signal filters. First, we high-pass filtered at 1Hz to remove the overall mean, and then low-pass filtered at 70Hz to remove high frequency noise. We also applied a notch filter to remove 60Hz noise.

Subject-generated EEG artifacts resulting from eye and muscle activity as well as global movements often have amplitudes that are much larger than signals originating from ordinary brain activity. Because artifact signals can seriously distort recorded brain signals, most EEG studies attempt to remove them during preprocessing. Unfortunately, there is no accepted standard practice for artifact removal and most published work augments automatic removal with manual identification.

A successful database retrieval framework requires a standardized approach to artifacts that does not include manual identification. The majority of the results in this paper were obtained for a database of data that did not have additional artifact removal beyond simple filtering (high, low, and notch). To test the effect of automated artifact removal, we also ran tests using the automated FASTER toolbox by Nolan et al. [19]. Internally, FASTER uses ICA to distinguish artifacts from brain signals. The effects of artifact removal are reported in Fig. 5.

C. Segmentation

The original EEG data consists of long continuous multi-channel signal recordings. For storage and retrieval, we segment these continuous recordings into epochs, which are equal-sized intervals time-locked to specific events. The BCI2000 datasets have 11 event types falling into three categories: baseline, task, and rest. The recording of baseline behavior is 1 minute, while each task is recorded for about 4.2 seconds immediately following the task trigger. The task behavior is followed by 4.1 seconds of rest behavior.

As indicated from Table III, we have tested the system with epoch lengths of 1, 2, 3, and 4 seconds. For segment lengths of 1 second, only the first second of signal in each epoch is retained. In case of baseline epochs, which have only one event for one minute recording, we assume there is a virtual event marker every 4 seconds and segment the baseline epoch based on these virtual markers. As a result, each subject has a total of 390 segments.

D. Feature extraction

As discussed in Section II, we test our framework using five different features: kurtosis, mobility, complexity, autoregressive coefficients, and the band intensity ratio of the alpha frequency band to the beta frequency band. All of the features were computed on each channel of each epoch independently.

E. Retrieval

To find the similar samples to a query sample, we calculate the distance between the query and test sample features using the sum of squared differences (SSD).

$$SSD(f_q, f_t) = \sum_{i=1}^L (f_q^i - f_t^i)^2$$

where f_q^i is i^{th} component of the query feature vector, f_t^i is the i^{th} component of the test feature vector, and L is the length of a feature vector.

After measuring distances of all possible pairs, samples are sorted according to their distance from the query sample. The system returns the ordered list of samples for each query.

F. Evaluation

We used three metrics: *MAP*, *ANMRR*, and *MNRO* to evaluate the performance. As explained in Section II, the range of the *MAP* metric is 0 to 1 for the worst to the best cases, while *ANMRR* and *MNRO* range from 1 to 0 for the worst to the best cases. To display these metrics on the same scale, we use $(1-ANMRR)$ and $(1-MNRO)$ on the graphs.

IV. Results

Using the proposed framework, we can evaluate the effects of various parameter choices under different retrieval scenarios. We restrict the target of retrieval to fixed length time slices of multichannel EEG.

For the experiments, two retrieval scenarios were used: same-subject retrieval and similar-subject retrieval. The goal of the same-subject retrieval is to find samples from the same subject. The goal of similar-subject retrieval is to find which subjects are most likely to be retrieved in response to queries from a particular subject. Similar subject retrieval can be used in a number of applications including propagating subject metadata and augmenting test datasets. Most of the results below are presented for same-subject retrieval. In this

retrieval task, each epoch is used as a potential query. The three metrics (*MAP*, *ANMRR*, and *MNRO*) are averages over all queries.

A. Effect of database size

We begin by testing the effect of database size using complexity as the feature. We suspect that because it is more difficult to find true samples in larger databases, the retrieval performance will degrade as the size of the database increases.

Fig. 2 compares the performance of same-subject retrieval for the three evaluation metrics when the complexity feature is used and the database size varies. Among 103 subjects, 10 to 100 subjects are randomly selected. Because each subject has 390 samples, 10 subjects are equivalent to 3,900 samples and 100 subjects are equivalent to 39,000 samples. The three metrics behave similarly, but *ANMRR* consistently reports better performance than the other two metrics. The results using the other features are similar.

Table IV shows the performance difference between a 10-subject database and a 100-subject database for the three metrics and different features. The *MNRO* metric appears to be more conservative and consistently shows a smaller difference in performance between 10 subjects and 100 subjects than the other two metrics. The *MNRO* metric will be used in the remainder of the paper.

B. Effect of feature selection

Fig. 3 shows the same-subject performance results of various features for different size databases. The autoregressive features, which capture signal dynamics, significantly outperform the other features in all cases.

AR feature vectors are of length 128 when a model order of 2 is used, while the other features generate feature vectors of length 64 for 64-channel EEG data when computed over the entire epoch. To make a fairer comparison, we equalized the feature length by computing the shorter features on two consecutive two-second intervals and concatenating the feature vectors of the two segments. These results are displayed in Fig. 3. We also evaluated the features computed on the full 4-second intervals and found that they did not change significantly from those presented here.

C. Effect of segment length

To measure the effect of segment or epoch size, we evaluated same-subject retrieval using the five different features extracted using different length segments. The same number of samples of each type was used as in the 4-second case. Fig. 4 shows the retrieval performance of each database when various lengths of segments are used. All features improve when the segment length is increased from 1 second to 4 seconds.

D. Effect of artifact removal

To assess the effect of artifact removal on retrieval performance, we measured performance using *MNRO* on the 50-subject database of 4-second segments for the same-subject retrieval task. Fig. 5 shows that artifact removal (or cleaning data) does not guarantee the

improvement of the retrieval performance. The retrieval performance of kurtosis and autoregressive (*AR*) decrease after artifact removal, while the performance of mobility is not affected by artifact removal. Only complexity and band intensity ratio (*BI Ratio*) appear to benefit of artifact removal.

These results are consistent regardless of the segment length. Fig. 6 shows that for *AR* the performance for the original data is always better than for the cleaned data in a 50-subject database regardless of the segment length.

E. Application: finding similar subjects

A potential application of EEG databases is to find similar subjects. We tested to see if the false positives retrieved for a particular subject tended to cluster in a small group of other subjects. We used the 103-subject database of 4-second segments and *AR* features.

Fig. 7 shows a typical result. The figure depicts a two-dimensional histogram of the top K query responses for Subject 7 using *AR* features and a segment length of 4 seconds, where K is twice the number of true same-subject samples. Because Subject 7 has 390 samples, the graph has 390 rows and K is 780. Each row shows the number of returned results from different subjects for a particular query example from Subject 7. The columns of the graph correspond to the counts of retrieved results for each subject in response to the individual Subject 7 queries. As Fig. 7 shows, the most popular subject in the top K samples is the query subject. However, Subjects 17, 32, and 66 also show a strong affinity for Subject 7 queries. A histogram of the counts averaged over the queries is shown at the top of the figure.

This result is consistent across query subjects. Each subject has a few subjects whose queries show a strong affinity for its queries. One can estimate a p-value for the likelihood that this number of matches could have occurred by random chance or simply calculate the ratios of average counts.

V. Discussion

This paper proposes a framework for content-based EEG retrieval (CBER). To verify the proposed framework, we implemented a simple EEG prototype database supporting query-by-example retrieval and conducted experiments with real EEG data using five different features and three evaluation metrics. These experiments ranged from basic feature comparison to the application of finding similar subjects. The results clearly show that content-based EEG retrieval is feasible and has many potential applications.

Although the numeric values of the chosen metrics are not close to one for same-subject retrieval, Fig. 7 demonstrates that if the query set is large enough (in this case twice the number of true positives), the majority of the correct samples will be among the retrieved results, as well as those from a small number of additional subjects. One way similar-subject retrieval might work in practice is for the user to present a collection of EEG epochs from a subject to the system. The system could evaluate similarity using the database, augmented by the subject's own samples. The results would provide a measure of how successful these

features were for self-retrieval. The best features for this subject could then be used to retrieve similar subjects in the database.

Autoregressive features appeared to be the most useful for the subject retrieval task and seemed to perform better for 4-second segments. Both kurtosis and autoregressive features performed better for data before artifacts were removed. In some ways this is not surprising, as these features are both effective in detecting artifacts [20]. It is possible that some of the effectiveness of *AR* features is their ability to capture eye activity. However, the FASTER automated artifact removal applied here is very aggressive in removing eye-activity as well as other activity regarded as noise. *AR* features still retained their advantage after artifact removal. It should be noted that the BCI2000 corpus was relatively free of artifacts and did not have a large amount of blink activity.

For test purposes, this work used an in-memory database. However, actual implementation on a larger scale requires a database management system (DBMS). We have implemented such a system on top of PostgreSQL [21]. While intrinsically a traditional relational database, PostgreSQL supports a number of features that facilitate its use as a CBER system. In particular, PostgreSQL supports the storage of arbitrary vectors as individual table elements and also supports K-nearest neighbor queries. This facility allows the pre-computation and storage of a large number of different features as well as information about how epochs were assigned to various types of classes using different classifiers. The database can then be mined to compare different features for different tasks on a fairly large scale.

We have run two retrieval scenarios: same-subject retrieval and similar-subject retrieval and demonstrated the potential of such approaches. Other retrieval tasks --- such as same task retrieval are more difficult, and our preliminary results with the five simple features presented in this paper showed large variations in performance among subjects and the particular tasks.

Ultimately, the success of retrieval will depend on the quality of the features relative to the retrieval task. Many approaches are possible for improving performance including different ways of combining features and post processing through secondary classifiers. Experiments to understand the brain and body during active behavior often provide additional measurements besides EEG, such as eye-tracking and movement sensors. Thus type of diverse information may serve to augment and annotate features for additional retrieval tasks.

Acknowledgments

Computational support for this project was provided by the Computational System Biology Core, funded by the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

References

1. Lew MS, Sebe N, Djeraba C, Jain R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans Multimedia Comput Commun Appl*. Feb; 2006 2(1):1–19.

2. Alto H, Rangayyan RM, Desautels JEL. Content-based retrieval and analysis of mammographic masses. *J Electron Imaging*. Apr; 2005 14(2):023016–1–023016–17.
3. Yang L, Jin R, Mummert L, Sukthankar R, Goode A, Zheng B, Hoi SCH, Satyanarayanan M. A boosting framework for visibility-preserving distance metric learning and its application to medical image retrieval. *IEEE Trans Pattern Anal Mach Intell*. Jan; 2010 32(1):30–44. [PubMed: 19926897]
4. Pawar, MM.; Pradhan, GN.; Zhang, K.; Prabhakaran, B. Content based querying and searching for 3D human motions. *Proceedings of the 14th international conference on Advances in multimedia modeling*; Berlin, Heidelberg. 2008. p. 446–455.
5. HBC, Pradhan G, Misra A, Prasanna SRM, Das RK, Sinha R. Multivariability speaker recognition database in Indian scenario. *Int J Speech Technol*. Dec; 2012 15(4):441–453.
6. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. Nov; 2012 41(D1):D8–D20. [PubMed: 23193264]
7. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Research*. Nov; 2012 41(D1):D36–D42. [PubMed: 23193287]
8. Lu H, Eng HL, Guan C, Plataniotis KN, Venetsanopoulos AN. Regularized Common Spatial Pattern With Aggregation for EEG Classification in Small-Sample Setting. *IEEE Transactions on Biomedical Engineering*. Dec; 2010 57(12):2936–2946. [PubMed: 20889425]
9. Ihle M, Feldwisch-Drentrup H, Teixeira CA, Witon A, Schelter B, Timmer J, Schulze-Bonhage A. EPILEPSIAE - a European epilepsy database. *Comput Methods Programs Biomed*. Jun; 2012 106(3):127–138. [PubMed: 20863589]
10. [Accessed: 21-Feb-2013] PostgreSQL: The world's most advanced open source database. [Online]. Available: <http://www.postgresql.org/>
11. Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*. Jun; 2004 51(6):1034–1043. [PubMed: 15188875]
12. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*. Jun; 2000 101(23):e215–e220. [PubMed: 10851218]
13. Hjorth B. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*. Sep; 1970 29(3):306–310. [PubMed: 4195653]
14. Bao, FS.; Gao, J-M.; Hu, J.; Lie, D.; Zhang, Y.; Oommen, KJ. Automated epilepsy diagnosis using interictal scalp EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2009. p. 6603–6607.
15. Bao FS, Liu X, Zhang C. PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction. *Computational Intelligence and Neuroscience*. 2011; 2011:1–7. [PubMed: 21837235]
16. Chatzichristofis SA, Iakovidou C, Boutalis YS, Angelopoulou E. Mean Normalized Retrieval Order (MNRO): a new content-based image retrieval performance measure. *Multimed Tools Appl*. Aug; 2012 :1–32.
17. Cieplinski, L. MPEG-7 Color Descriptors and Their Applications. In: Skarbek, W., editor. *Computer Analysis of Images and Patterns*. Springer; Berlin Heidelberg: 2001. p. 11–20.
18. Deselaers T, Keysers D, Ney H. Features for image retrieval: an experimental comparison. *Inf Retrieval*. Apr; 2008 11(2):77–107.
19. Nolan H, Whelan R, Reilly RB. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*. Sep; 2010 192(1):152–162. [PubMed: 20654646]
20. Lawhern V, Hairston WD, McDowell K, Westerfield M, Robbins K. Detection and classification of subject-generated artifacts in EEG signals using autoregressive models. *J Neurosci Methods*. Jul; 2012 208(2):181–189. [PubMed: 22634706]
21. Hossain, A.; Cockfield, J.; Robbins, K. MoBBED (Mobile Brain-Body-Environment Decision-making) Data Infrastructure (White Paper). University of Texas; San Antonio: 2012.

Figure 1.
The basic workflow

Figure 2.
Performance for complexity using different size databases.

Figure 3.
Retrieval performance of various features and database sizes

Figure 4.
Segment length vs. retrieval performance (50 subjects)

Figure 5.
Original sample vs. artifact removed sample

Figure 6.

Original sample vs. artifact removed sample for the autoregressive features using different segment lengths

Figure 7.

Distribution of retrieved subjects (Query subject: 7, excluded subjects: 88, 89, 92, 100, 104, 106)

TABLE I

THE LIST OF FEATURES

Category	Name	Description	Feature size ^a
Statistical	Kurtosis (f_K)	Measure of outlier samples (weight of distribution tails)	L
Temporal	Mobility (f_M)	Mean frequency of spectral density	L
	Complexity (f_C)	Measure of excessive details	L
	Autoregressive (f_A)	Time dynamics of each channel	$P \times L$ (P : model order)
Spectral	Band Intensity Ratio (f_B)	Ratio of spectral intensities in specific frequency ranges	L

^aFeatures are generated for a windowed signal represented by an $L \times N$ array, where L is the number of channels and N is the number of samples in a window. An input dataset of size $L \times N \times K$ produces K feature vectors.

TABLE II

COMMON NOTATIONS

Symbol	Values	Meaning
X	$[x_1, x_2, \dots, x_N]$	Time series values (i.e., ordered samples)
D	$[d_1, d_2, \dots, d_{N-1}]$, where $d_i = x_i - x_{i-1}$	First difference of X
S	$[s_1, s_2, \dots, s_{N-2}]$, where $s_i = d_i - d_{i-1}$	Second difference of X
μ_X	$\frac{1}{N} \sum_{i=1}^N x_i$	Mean of X
M_4	$\frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^4$	The fourth moment about the mean for X
σ_X	$\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2}$	Standard deviation of X
σ_D	$\sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (d_i - \mu_D)^2}$	Standard deviation of first difference D of X
σ_S	$\sqrt{\frac{1}{N-2} \sum_{i=1}^{N-2} (s_i - \mu_S)^2}$	Standard deviation of second difference S of X

TABLE III

PARAMETERS USED TO ILLUSTRATE FRAMEWORK.

Stage	Values
EEG data	37,286 epochs from 103 subjects
Noise and artifact removal	High-pass (1Hz), Low-pass (70Hz), Notch (60Hz) Remove eye artifacts, Remove outlier ICs
Segmentation	Lengths of 1, 2, 3 or 4 seconds
Feature extraction	Kurtosis, Mobility, Complexity, Autoregressive, or Band Intensity Ratio
Retrieval	Sum of squared differences
Evaluation	<i>MAP, ANMRR, MNRO</i>

TABLE IV**AVERAGE PERFORMANCE DEGRADATION^b**

	MAP	ANMRR	MNRO
Kurtosis	0.21	0.24	0.19
Mobility	0.37	0.34	0.28
Complexity	0.32	0.32	0.26
Autoregressive ($p = 2$)	0.36	0.32	0.26
Band Intensity Ratio	0.31	0.35	0.27

^b Performance difference between database sizes of 10 and 100 subjects.