2013 IEEE International Conference on Systems, Man, and Cybernetics

Evaluation of Different Onscreen Keyboard Layouts using EEG signals

Arijit Sinharay, Debatri Chatterjee, Aniruddha Sinha
Tata Consultancy Services Ltd.
Innovation Lab, Kolkata
arijit.sinharay@tcs.com,debatri.chatterjee@tcs.com,aniruddha.s@tcs.com

***Abstract*—The paper aims at evaluation of different onscreen keyboard layouts based on the biological responses of the users. The signal used for the said purpose is Electroencephalogram acquired by low cost neuro-headset from Emotiv. We propose to use human cognition as the fundamental feature to discriminate between user-friendly vs. cumbersome onscreen layout designs. To validate our observations we compared our results with bench marked data based on user study and KLM–GOMS model. A classifier is first trained for high and low cognition tasks based on well-established cognitive tests (e.g Stroop test) and then this classifier is used to report the cognition class for a particular onscreen layout. A high cognition load class indicates complexity in the layout design whereas a low cognition output indicates the layout to be user friendly. Present evaluation methods like user study or KLM-GOMS based model, serves as an indirect measure of goodness of layout designs. In contrast, our approach has a unique advantage as this is a direct measure of human's biological response subjected to stimuli (in our case onscreen keyboard layouts) hence more reliable.**

Keywords-component; KLM-GOMS, Cognitive load, keyboard layout evaluation, Electroencephalogram

I. INTRODUCTION

Nowadays internet is the vast source of information and entertainment. Moreover, the convergence of different access technologies and access devices are also gaining popularity day by day. Multiple applications like voice call, video call, television programs, internet etc. can be accessed using a single access device. It reduces the overall cost of the home entertainment system. Some of such access devices are personal computer (PC), laptop, television, set-top box etc. In almost all developing countries television penetration is much higher compared to that of PC. Hence television can be successfully used as the single access device for all converging applications [1], [2]. For this to happen, there is a need to enable the users to input text to the system, in a convenient manner. Traditional keyboards increase the overall cost of the system and is also very uncomfortable to use, as the user is sitting at a distance of 8-10 feet from the display device (e.g. television). Thus an onscreen keyboard operated through remote control is perceived to be one of the best possible options.

This brings out the importance of an on-screen keyboard layout which can be used for rapid and accurate text entry. Different onscreen keyboard layouts are being used for text entry but they are mostly restricted to multi-tap keyboards or QWERTY keyboard configurations [3]-[4].

Majority of these works lack any systematic means to evaluate their merits based on a metric involving the cognitive process of the subjects which is more direct measure of user friendliness of a particular layout. KLM-GOMS model serves as a theoretical measure to evaluate the degree of complexity of each layout in terms of text entry mechanism [8].

In this paper we have considered two onscreen keyboard layouts. A two-level Hierarchical organization of symbols and characters for onscreen keyboard is described in [5]. To evaluate and monitor the layout, a user study was conducted and the findings are given in [6]. Some modifications were done based on the results of the user study. KLM-GOMS model was used to theoretically evaluate the layout [7]. The results of KLM-GOMS model matched well with that obtained through user study.

In this paper, we attempt to experimentally validate the previous findings by measuring biometric responses of the subjects through electroencephalogram (EEG) signal analysis. The analysis reveals the imparted cognitive loads on the subjects while performing a particular task. Once validated, our technique can be used as a direct measure of user friendliness for any layouts.

The Cognitive load is defined as the amount of load imposed on the working memory by any task. It mostly depends on the working memory capacity [9]. An optimum level of cognitive load should be applied on brain to attain best performance level. There are three different types of cognitive load: i) Intrinsic ii) Extraneous and iii) Germane [10], each having an effect on learning and decision making tasks. The first one refers to the complexity of the task itself. The second one is the additional load on the user imposed by the method of representation of the information. The last one is the cognitive load generated during the learning phase of any new technology or schema. The layout evaluation is mostly related to extraneous load.

The choice of EEG to measure the cognitive load lies in its inexpensiveness, easy wear-ability, direct accessibility of scalp potentials (evoked during cognition) and excellent time

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.88

resolution when compared with other biometrical measurements. Alternative approaches to measure cognition involves measuring pupil dilation, galvanic skin response [12, 13], functional magnetic resonance imaging (fMRI) [17], functional near infrared spectroscope (fNIRs) [22], positron emission tomography (PET) [23] etc.

In the present work, we have used a low cost 14-lead EEG device from Emotiv (www.emotiv.com). The EEG data is acquired while the subject is typing on an onscreen layout of Television using remote. The EEG signals are processed in a window by window basis and for each window corresponding cognition class is reported. We restricted ourselves to two class classification problems; hence, either high or low cognition class is reported per window. Subjects are given two different onscreen keyboard layouts, consisting of our proposed Hierarchical onscreen layout (Fig. 1) and traditional QWERTY layout (Fig. 2) and are asked to type same phrases using both. Captured EEG signals are processed and analyzed to reveal the corresponding cognitive load class.

Prior to report the cognition class of a layout, we calibrate the classifier for high and low cognition classes based on the training sets that uses bench mark data for two different cognition levels.

The user studies conducted in [6] showed that the Hierarchical layout was easier to use compared to traditional QWERTY keyboard layout which is in accordance with our findings. Subjective measures of cognitive load like NASA-TLX test [11], user studies and post-experimental surveys have been included in our work to enhance reliability.

The paper is organized as follows. Section II provides the overview of two different onscreen layouts and method of operating those. Section III gives the details of EEG data acquisition and analysis tools. Section IV gives methodology of EEG measurement and processing. Details of the experimental procedures are given in section V, followed by results in section VI. Finally, the paper is concluded in section VII.

II. ONSCREEN KEYBOARD LAYOUTS

A Hierarchical onscreen keyboard has been proposed in [5] and [6]. The characters in the on-screen are organized into blocks of up to 4 characters or symbols, or character-sets.

Fig 1. Hierarchical onscreen keyboard layout

The traditional onscreen QWERTY keyboard layout is shown in Fig. 2.

Fig. 2. QWERTY keyboard layout

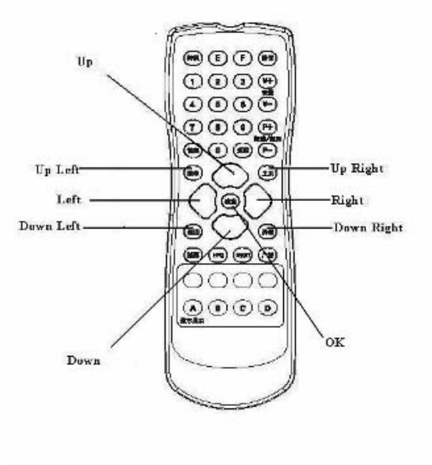

Fig. 3. Remote control for onscreen keyboard

Both the keyboards can be operated using the remote shown in Fig. 3. The navigational keys (Up, Down, Left and Right keys) of Fig. 3 are used to move across the key-blocks in both horizontal and vertical directions. Selection keys (Up Left key, Up Right key, Down Left key, Down Right key and OK key) are used to select a particular character within a block. The switching between lower case letter screen, upper case letter screen, symbol screen and any other types of screens can also be achieved through specially assigned hot keys in the remote control. This particular keyboard layout was evaluated both theoretically and through user study [6], [7]. Based on the reports obtained through user study, some important modifications have been adopted in the proposed layout.

For theoretical performance evaluation of this onscreen layout, KLM-GOMS model has been used. In order to model the remote based operations, the standard KLM operator set has been extended. The extended KLM-GOMS model has been discussed in [24].

III. EEG DATA ACQUISITION AND ANALYSIS TOOL

For recording the EEG signals, while executing a task, we have used 14-channel Emotiv neuro-headset. The sampling frequency of the device is set at 128Hz. The Emotiv SDK gives us the access to the raw EEG data which are pre-filtered for common noise suppression using notch filter at the power line frequency. The Emotive EEG system uses standard 10-20 electrode configuration about channel fixing. It also has two reference channels (DRL, CMS) at P3/P4 locations, offering optimal positioning for good spatial resolution. The headset is shown in Fig.4

Fig. 4. Emotiv neuro headset

Instruction for each task is presented to the user through a computer screen and time synchronized EEG recording was done for post processing. For data entry through onscreen keyboard we have used a SONY Bravia Television set. We made each subject to try out the layout prior to the start of the experiment. This eliminates the Germane load on the subjects. We have used MATLAB for all our post processing needs.

IV. METHODOLOGY

The evaluation methodology can be broadly divided into two steps:1)training a standard classifier for high and low cognitive loads using bench-marked data and standard psychometric tests, 2) using this classifier to report cognition class for the onscreen keyboard layouts (QWERTY and Hierarchical).

We now address pre-processing and feature extraction of the EEG data. EEG data is processed in a window by window basis and the length of a time window is experimentally chosen for best performance. We will describe the window selection shortly.

The raw EEG scalp signals (in a given window) have poor spatial resolution. Thus we have used common spatial pattern (CSP) filtering to find out the most discriminative components of the acquired EEG signals. CSP aims at learning spatial filters which maximize the variance of band pass filtered EEG signals from one class while minimizing their variance from other class [19]. On the filtered data, we calculated log variance of two most significant CSP channels (chosen according to the most significant Eigen values, in other words the principal components). These two channels construct the feature space. Thus, the feature vectors are two dimensional in nature. Moreover, features are normalized before any training or classification is done.

If $x_i(t)$ is the filtered output from ith CSP channel for an analysis window then the corresponding feature is defined as:

$$f_i = \log(Variance\ (x_i(t)) \qquad (1)$$

where i=1,2 for two channels.

Features are normalized as:

$$f_i = \frac{f^i - f^i_{\min}}{f^i_{\max} - f^i_{\min}} \qquad (2)$$

Where, $f^i_{\min}$ and $f^i_{\max}$ represents the minimum and maximum values of the feature on the ith channel for all the windows of a given trial. In this paper, we refer trial as the complete duration of a task performed by a subject.

Finally, the normalized features construct the feature vector as:

$$\vec{f} = [f_1, f_2] \qquad (3)$$

Where, f_1 is obtained from 1st CSP channel and f_2 is obtained from 2nd CSP channel.

The feature vectors derived from the benchmark data are used for creating a training model for a two level linear SVM classifier using radial basis function (RBF) kernel. The trained model is later used to classify the test data set into two cognitive load classes.

We further define Cognition Score (CS) based on the percentage of windows reported 'High' for an entire trial. For example, if a particular trial consists of N analysis windows and if M windows are reported as 'High', then CS for that trial is defined as:

$$CS = \frac{M}{N} \times 100\% \qquad (4)$$

Thus higher CS means higher cognition load imparted on the subject for that particular trial.

The overall process for measuring cognitive load is depicted in Fig. 5.

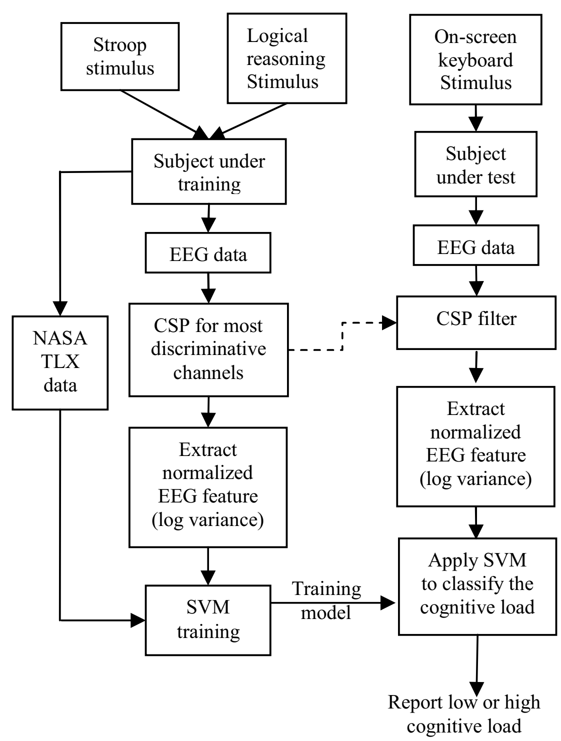

Fig. 5. Overall process for measuring the cognitive load

A. Finalizing the window size

Since our analysis is window based, we first tried to find a window size that is best for classifying the Stroop data set .This widow size is then fixed and applied to all the trials. This is to be noted that while doing windowing, we apply non-overlapping sliding windows as depicted in Fig.6.

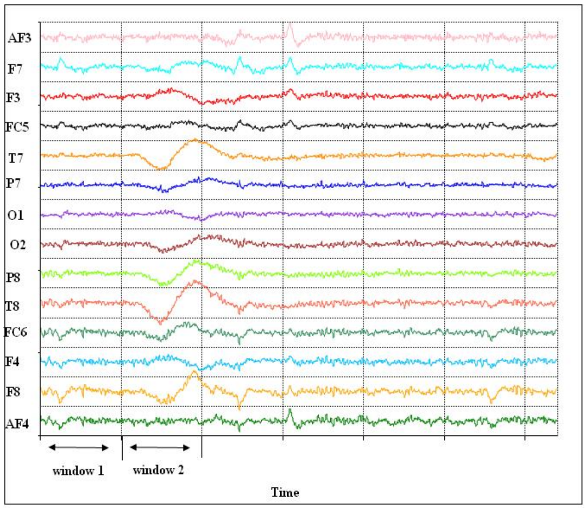

Fig. 6. Raw EEG signals for window based analysis

To identify a window size, we calculate CS for each subject doing a high and a low cognition operation based on Stroop test. For a particular subject and for a particular window size w, we define:

$CS^i_{H,w}$ is the cognitive score for high trials

and $CS^i_{L,w}$ is the cognitive score for low trials

Discriminative Index (DI) for that particular window is defined as:

$$DI_w{}^i = CS^i_{H,w} - CS^i_{L,w} \quad (5)$$

We find the window size that maximizes the DI using all user data. Mathematically it can be represented as:

$$w_{opt} = \underset{w \in [j | l1 < j < l2]}{\arg\max} \left(\sum_{i=1}^{n} DI^i_w \right) \quad (6)$$

Where w_{opt} represents the optimal window size, n represents the total number of subjects, $l1$ represents lower end of window length and $l2$ represents higher end of window length.

B. Training phase with Stroop test (Benchmark data set I)

Two sets of stimuli were designed based on standard Stroop test [20], [21].

In first set (Fig. 7), the name of the colour was printed in a colour denoted by the name. The users were asked to read the colours while wearing the EEG headset.

In the second set (Fig. 8) the name was printed in a colour not denoted by the name. In this case the users were asked to read the colour of the text instead of name of the colour.

It is well established fact that the first set imparts lower cognitive load as compared to the second set.

Fig. 7. Set for low cognitive load

Fig. 8. Set for high cognitive load

C. Training phase using designed stimulus (benchmark data set II)

Another set of experiments were designed based on mathematical calculations and logical reasoning tasks for two high and low levels of cognitive load. Task difficulty level was varied by number of digits of calculation and complexity of logical reasoning. Two such experimental slides are given in Fig. 9 and Fig. 10. Fig.9 corresponds to low cognitive load whereas Fig.10 corresponds to high load.

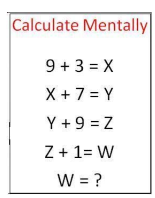

Fig. 9. Experiment for low load

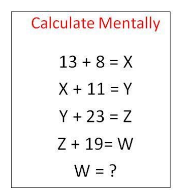

Fig. 10. Experiment for high load

D. Validation of training data

On completion of each trial, subjects were asked to rate the cognitive loads of that particular task using NASA task load index [25] which assesses the users work load in six dimensions such as mental, physical and temporal demands, own performance, effort and frustration. Considering all the dimensions, the cognitive load level for a particular task was finally evaluated. The results obtained from all subjects were considered to finally evaluate the cognitive load of a particular task. This exercise was done to validate our designs of low and high cognitive tasks, as the cognitive load imposed by a particular task might be different for different persons.

Participant's opinion (from NASA TLX) about complexity of the designed stimuli matched well with our design considerations. Cognition load level of Stroop test is also well known. Hence these two were treated as the ground truth for cognitive load classification and were used to train a classifier.

E. Testing phase with onscreen keyboard

In the testing phase, the users were asked to type some phrases using both QWERTY (Fig. 2) and the Hierarchical layout (Fig. 1). The texts were selected from MacKenzie's standard phrase sets [18]. The average length of phrases selected was 25 characters. To study the brain activities during the task corresponding EEG signals were acquired.

V. EXPERIMENTS

Three separate experiments are conducted for evaluation of onscreen keyboard layouts. This is to be noted that we only invited male participants as it was found to be difficult to get good electrode contact for women having long hair.

A. Experiment 1

A group of 20 subjects are selected in the age group of 25-30 years. This experiment is carried out to get one sets of training data (benchmark dataset I). Subjects are asked to take standard Stroop test and the corresponding brain signals are recorded.

B. Experiment 2

Subjects participated in experiment 1 are also asked to take part in designed stimulus test (benchmark dataset II) as described in section IV-C. Each participant completed both the sessions of low and high cognitive tasks with a short break of 30 seconds between each session.

After completing each experiment (experiment 1 and experiment 2), the subjects are requested to log their experiences using NASA TLX index.

C. Experiment 3

A group of 10 subjects are selected from the participants who took part in experiment 1 and 2. In addition we also selected 5 subjects who did not participated in experiments 1 and 2. All of them are given onscreen layouts shown in Fig.1 and Fig.2 and are asked to type few phrases using those. While typing, the EEG signals are recorded.

VI. RESULTS

Here we present the results of analysis of window length and that of three experiments as mentioned in the previous section.

A. Analysis of window length

An analysis is done to find the effect of window length and to choose an optimal window size. In order to do this, the window length is varied from 5 seconds to 20 seconds in steps of 5 seconds for benchmark training data set I. DI has been calculated and the best window size is found using (6). It is found that window size of 10 seconds is optimal for majority of subjects. Sample results for 5 subjects are shown Fig. 11.

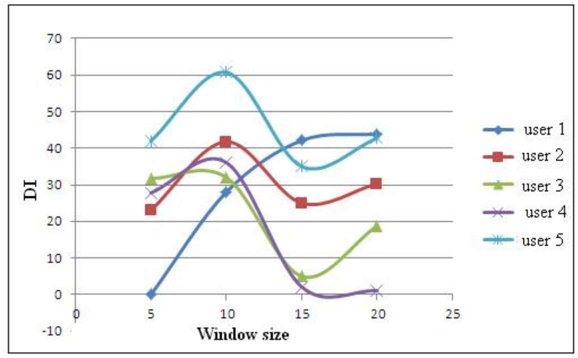

Fig. 11. Effect of window length is shown. Average DI is highest around window size of 10 seconds

B. Results of experiment 1

It is found that the cognitive load measured from EEG signals acquired during experiment 1 can be distinctly classified into two groups. The results are shown in the Fig. 12. The circles 'o' corresponds to the data points for high cognitive load and the '+' signs corresponds to that of low cognitive load. The result obtained are consistent with Stroop test, i.e, first stroop set data mostly got classified as low cognition class whereas the second set mostly falls under high cognition class.

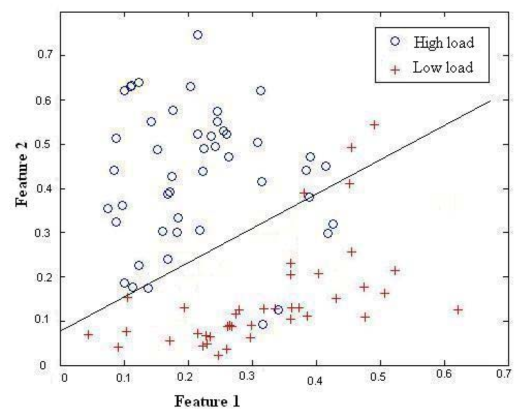

Fig. 12. Plot for Stroop data (set I) is shown. Two distint clusters corersponding to high and low cognitive loads are obserevd

C. Results of experiment 2

This experiment consisted of two different set of tasks as discussed in section V-C. The results obtained for these two sets of tasks are found to be discriminative in nature and in accordance with the designed stimulus. The result for a particular subject is shown in Fig. 13.

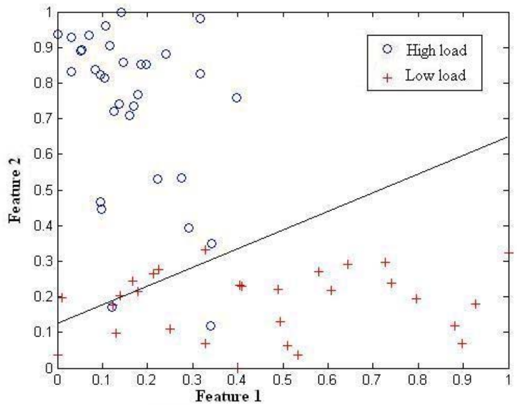

Fig. 13. Plot for the designed stimulus (set II) is shown. Two distinct clusters corresponding to low and high cognitive loads are observed.

D. *Training the classifier*

The data sets (set I and set II) from both the experiments (i.e. experiments 1 and 2) for all subjects are used as the training data set for generating the SVM model. This model is then used to classify the test data which consist of tasks related to input text for different onscreen layouts.

E. *Results of experiment 3*

The features extracted from the captured EEG data, while a subject is using a particular type of on-screen keyboard, are then fed to a previously trained SVM classifier to determine the cognition class for each window. Fig. 14 shows the classification results of EEG signals for both the layouts for a particular subject. The partition is obtained by training the SVM classifier (Fig 12 and Fig 13). Our result clearly indicates that Hierarchical layout is easier than QWERTY layout as it falls under low load section. This is also in accordance with the results obtained from user study and KLM-GOMS. Finally we have calculated CS for each trial for all subjects.

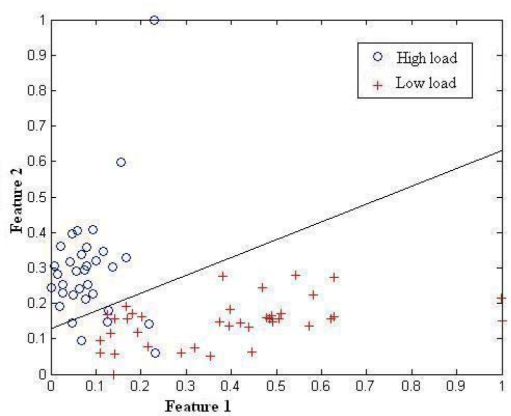

Fig. 14. Two Onscreen keyboard data for a single subject is shown. Two distinct cluster can be found indicating corresponding high and low cognitive load

The summary of CS values for two categories of subjects (who were part of the training set collection process, and subjects who did not participated in training data collection) is given in Table I and Table II. It can be seen from Table I and Table II that the CS for QWERTY keyboard is higher than that of the Hierarchical keyboard for all the subjects. However, the average DI for two types of keyboard is lower in Table II. This indicates that if the subjects under test participate during the generation of the training model, then the classification yields better results. Fig.15 and Fig.16 are pictorial representation of Table I and Table II.

Experimentally it was found that searching a character and then moving focus to that character was indeed difficult for onscreen QWERTY layout than the layout shown in Fig. 1.

TABLE I. COMPARISON OF THE COGNITIVE SCORE FOR TWO TYPES OF KEYBOARDS (SUBJECTS WERE PART OF THE TRAINING SET)

Subjects	CS for QWERTY keyboard	CS for Hierarchical keyboard	Average DI
1	66	38	38.13
2	81.85	40.1	
3	96.1	65.1	
4	84.2	48.12	
5	96	25.3	
6	85	25.7	
7	96.8	62.1	
8	72.5	51.2	
9	90.2	65.1	
10	72.1	40.1	

TABLE II. COMPARISON OF THE COGNITIVE SCORE FOR TWO TYPES OF KEYBOARDS (SUBJECTS WERE NOT PART OF THE TRAINING SET)

Subjects	CS for QWERTY keyboard	CS for Hierarchical keyboard	Average DI
1	86.42	60	33.22
2	91.2	63	
3	85	69.2	
4	84.2	59.2	
5	96	25.3	

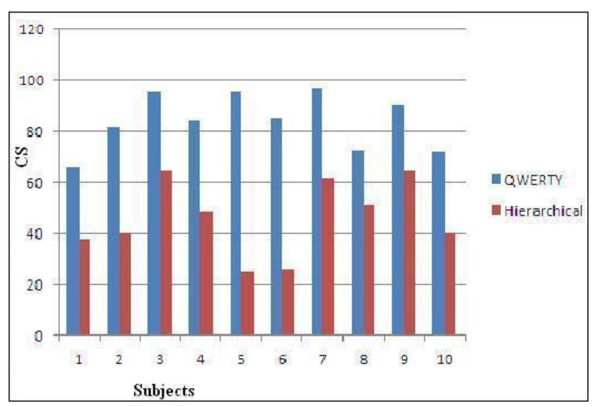

Fig. 15. CSs for 10 subject (who took part in training phase) for two different keyboard layouts have been shown. It clearly shows that Qwerty imparts higher cognitive load than the Hierarchical layout as CS value is higher for the Qwerty layout

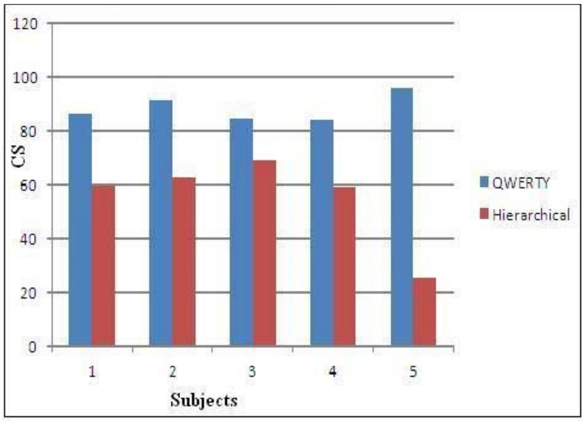

Fig. 16. CS for subjects not participated in training. The average DI is less here as the classifier was not particularly trained for these subjects

VII. CONCLUSIONS

Our suggestion of using cognitive load as an important proposition to evaluate the user friendliness of a particular onscreen layout seems promising based on the experimental results. Moreover, measuring cognitive loads through low cost EEG system is also suited for all practical purposes. Our approach correctly classified different on-screen keyboard layouts that has been already evaluated by user study and supported by KLM-GOMS model. The proposed approach has an advantage that it directly measures brain signals which in turn indicates subject's perceived difficulty to use a particular type of keyboard. Hence, our approach proposes a novel way of evaluating different onscreen keyboard layouts for application in effective User Interface design. In future this work can also be extended for multiclass classification among different layouts rather than just high or low as reported in the present work. Moreover, single lead Neurosky EEG device could also be tried for data acquisition as it is more compact, user friendly and much cheaper.

ACKNOWLEDGEMENTS

We would like to thank all the subjects participated in the user study and also the group of people who appeared for the experiments conducted for recording the EEG signals.

REFERENCES

[1] World Internet Usage Statistics News and World Population Stats.http://www.internetworldstats.com/stats.htm.

[2] ITU report baffled over RP's high mobile phone, TV penetration. http://technews.com.ph/?p=1627.

[3] Iatrino and S. Modeo, "Text editing in digital terrestrial television: a comparison of three interfaces". In Proceedings of EuroITV'06, Athens, Greece,2006.

[4] M. Ingmarsson, D. Dinka, S. Zhai, "TNT-A numeric keypad based text input method" , Proc. CHI 2004: ACM Conference on Human Factors in Computing Systems. Vienna, Austria. CHI Letters 6(1), 639-646, ACM Press.

[5] A. Pal, C. Bhaumik, D. Kar, S. Ghoshdastidar, J. Shukla, "A novel on-screen keyboard for hierarchical navigation with reduced number of key strokes", IEEE International Conference on Systems, Man and Cybernetics, 2009. SMC 2009.

[6] A. Pal, D. Chatterjee, D. Kar, " Evaluation and Improvements of on-screen keyboard for TV and set -top box", 15th IEEE International symposium on Consumer Electronics, June 2011, Singapore.

[7] A. Basu, D. Chatterjee, A. Pal, "Extended KLM-GOMS modeling and Evaluation of onscreen keyboard layouts for television based internet applications", accepted in 4th international conference on Human Computer Interaction , Pune, India, April 2012.

[8] S. K. Card, T. P. Moran, A. Newell, "The Keystroke-Level Model for User Performance Time with Interactive Systems". *Comm. ACM 23, 7.* 396-410. 1980

[9] J. Sweller, The Cambridge handbook of multimedia learning. Cambridge University Press, 2005, ch. Implications of cognitive load theory for multimedia learning, pp. 19–30.

[10] P. Chandler, J. Sweller "Cognitive load theory and the format of instruction. Cognition and Instruction 8 (1991),pp. 293–332.

[11] J. M. Hitt, J. P. Kring, E. Daskarolise, C. Morris , M. Mouloua.: "Assessing mental workload with subjective measures: An nalytical review of the nasa-tlx index since its inception". Human Factors and Ergonimics Society Annual Meeting 43 (1999), 1404–1404.

[12] Y. Shi, N. Ruiz , R. Taib, E. Choi, F. Chen: "Galvanic skin response (gsr) as an index of cognitive load." In CHI '07 extended abstracts (2007), ACM, pp. 2651–2656.

[13] J. Klingner, B. Tversky, P. Hanrahan: "Effects of visual and verbal presentation on cognitive load in vigilance, memory and arithmetic tasks". Journal of Psyhophysiology 48, 3 (2011),323–332.

[14] W. Klimesch, "EEG alphs and Theta oscillations reflect cognitive and memory performance: a review and analysis" in Brain Research Reviews 29, 1999.

[15] T. Lan, A. Adami, M. Pavel , "Estimating Cognitive State Using EEG signals" in Biomedical Engineering Department, Oregon Health & Science University in EUSIPCO 2005.

[16] P. Antonenko, F. Paas, R. Grabner, "Using Electroencephalography to Measure Cognitive Load" in Educ Psychol Rev (2010).

[17] F. Lotte, J. Van Dille Dirk, L. Koole, "Tuning down the emotional brain: An fMRI study of the effects of cognitive load on the processing of affective images," NeuroImage 2009.

[18] S. I. MacKenzie, and W. R. Soukoreff, "Phrase sets for evaluating text entry techniques," in CHI '03 extended abstracts, Lauderdale, USA, 754-755, 2003.

[19] F.Lotte, C.Guan, "Regularizing common spatial patterns to improve BCI designs: Unified Theory and new Algorithms", in IEEE transaction on bio-medical Engineering vol. x, pages 355-362.

[20] Alverage, A. Julie, E. Emory, "Executive function and the Frontal Lobes: A Meta-Analytic Review" in Neoropsychology Review, vol 6, No.I, March 2006, DOI:10.1007/s11065-006-9002-x

[21] J. Demakis, George, "Frontal Lobe Damage and Tests of Executive Processing: A Meta-Analysis of the Category Test, Stroop Test, and Trail-Making Test" in Journal of Clinical and Experimental Neuropsychology, Vol 26, Issue3, 2004

[22] E. T. Solovey, P. Schermerhorn, M. Scheutz, A. Sassaroli, S. Fantini, R.J.K. Jacob, "Brainput: Enhancing Interactive Systems with Streaming fNIRS Brain Input", in CHI12, May 510, 2012, Austin, Texas, USA.

[23] E. E. Smith, J. Jonides, R.A. Koeppe, "Dissociating verbal and spatial working memory using PET in Cerebral Cortex",1996.

[24] D. Chatterjee, A. Sinha, A. Pal, A. Basu, "An Iterative Methodology to Improve TV Onscreen Keyboard Layout Design through Evaluation of User Studies", Scientific and academic publishing journal, Advances in computing, doi: 10.5923/j.ac.20120205.03, pages 81-91.

[25] http://humanfactors.arc.nasa.gov/groups/TLX/downloads/NASA-TLXChapter.pdf