Annual Report for AOARD Grant FA2386-12-1-4049

Robust Multimodal Cognitive Load Measurement (RMCLM)

March 26, 2013

Name of Principal Investigators (PI and Co-PIs) : Fang Chen
- e-mail address : fang.chen@nicta.com.au
- Institution : National ICT Australia (NICTA)
- Mailing Address : Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia
- Phone : +61 2 9376 2101
- Fax : +61 2 9376 2025

Period of Performance: March/26/2012 – March/25/2013

Abstract: This report summarizes the important research activities, study results and research accomplishments out of the RMCLM project in the past year period. The objective of this project includes research of the fundamental issues related to the use of multiple input modalities and their fusion to enable robust and automatic cognitive load measurement (CLM) in the real world. Firstly, we carried out a further literature review on physiological measures of cognitive workload to include the recent advances of physiological measures of cognitive workload. In the meantime, we examined the use of various features (e.g. spectral and approximate entropies, wavelet-based complexity measures, correlation dimension, Hurst exponent) of electroencephalogram (EEG) signals to evaluate changes in working memory load during the performance of a cognitive task with varying difficulty/load levels. Eye based CLM was also studied. Three types of eye activity were investigated: pupillary response, blink, and eye movement (fixation and saccade). We further investigated the linguistic feature based CLM in this study and analyzed novel linguistic features as potential indices of cognitive load. All together, we had carried out CLM study of three unobtrusive modalities, namely EEG, eye activity, and linguistic feature based CLM, in the past year period.

List of Publications

a) Papers published in peer-reviewed journals

[1] Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, A., Taib, R., Yin, B. and Wang, Y., "**Multimodal Behaviour and Interaction as Indicators of Cognitive Load**", ACM Transactions on Interactive Intelligent Systems, vol. 2, no. 4, article 22, December 2012.

[2] Khawaja, M. A., Chen, F., Marcus, N., "**Analysis of Collaborative Communication for Linguistic Cues of Cognitive Load**", International Journal of Human Factors and Ergonomic Society, vol. 54, no 4. pp 518-529, August 2012.

[3] Chen, S., Epps, J., "**Automatic Classification of Eye Activity for Cognitive Load Measurement with Emotion Interference**," Computer Methods and Programs in Biomedicine, 2012.

b) Papers published in peer-reviewed conference proceedings:

[4] Zarjam, P., Epps, J., Chen, F., and Lovell, N. H., "**Classification of Working Memory Load Using Wavelet Complexity Features of EEG Signals**." Lecture Notes in Computer Science, vol. 7664, pp. 692–699, Nov. 2012, Springer-Verlag Berlin.

[5] Zarjam, P., Epps, J., and Lovell, N. H., "**Characterizing mental load in an arithmetic task using entropy-based features**." Proc. of the 11th International Conference on Information Science, Signal Processing and their applications (ISSPA'2012), pp. 199 – 204, 2012.

[6] Zarjam, P., Epps, J., Lovell, N. H., and Chen, F., "**Characterization of Memory Load in an Arithmetic Task using Non-Linear Analysis of EEG Signals**", Proc. of the 34th IEEE Engineering in Medicine and Biology Conference (EMBC'2012), pp. 3519-3522, California, USA, 2012.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
05 APR 2013	**Annual**	**26-03-2012 to 15-03-2013**

4. TITLE AND SUBTITLE
Robust Multimodal Cognitive Load Measurement

5a. CONTRACT NUMBER
FA23861214049

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Fang Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
National ICT Australia (NICTA),Level 5, 13 Garden Street,Eveleigh, NSW Sydney 2015,Australia,NA,NA

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD, UNIT 45002, APO, AP, 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AOARD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD-124049

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report summarizes the important research activities, study results and research accomplishments out of the RMCLM project in the past year period. The objective of this project includes research of the fundamental issues related to the use of multiple input modalities and their fusion to enable robust and automatic cognitive load measurement (CLM) in the real world. Firstly, we carried out a further literature review on physiological measures of cognitive workload to include the recent advances of physiological measures of cognitive workload. In the meantime, we examined the use of various features (e.g. spectral and approximate entropies, wavelet-based complexity measures, correlation dimension, Hurst exponent) of electroencephalogram (EEG) signals to evaluate changes in working memory load during the performance of a cognitive task with varying difficulty/load levels. Eye based CLM was also studied. Three types of eye activity were investigated: pupillary response, blink, and eye movement (fixation and saccade). We further investigated the linguistic feature based CLM in this study and analyzed novel linguistic features as potential indices of cognitive load. All together, we had carried out CLM study of three unobtrusive modalities, namely EEG, eye activity, and linguistic feature based CLM, in the past year period.

15. SUBJECT TERMS
Brain Science and Engineering, Computer and User Interface, Congnitive Modeling, Cognitive Modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	

Contents

1. Introduction

Historically, cognitive load has been measured using subjective self-rating scales (e.g. NASA TLX) and by performance scores, however these methods are post-hoc, are not feasible in all applications and are either subjective (self-rating) or not indicative of spare mental capacity (performance). There is a need for objective measures of cognitive load that are non-intrusive and objective, and have the potential to be determined in real time, i.e. measured continuously through the task.

This project has focused on three main modalities, namely electroencephalogram (EEG), eye activity, and linguistic features, for automatic cognitive load measurement.

2. Updated Literature Review

We carried out a further literature review on physiological measures of cognitive workload. The further investigation was focused on the recent advances of physiological measures of cognitive workload: eye movement, skin temperature, linguistic features, speech signals, EEG, Galvanic Skin Response (GSR), and pen input features.

3. EEG Based CLM

EEG is a noninvasive neuroimaging technique widely used for measuring cognitive workload, which offers high temporal resolution, ease of use, and a comparably low cost. We investigated different analysis method of electroencephalogram (EEG) signals to examine changes in working memory load during the performance of a cognitive task with varying difficulty levels.

3.1 Non-Linear Analysis of EEG Signals for CLM

Experiment: EEG signals were recorded during an arithmetic task while the induced load was varying in seven levels from very easy to extremely difficult. We studied six male participants, between the ages of 24-30 years. They were right-handed and had normal or corrected to normal eyesight and gave written informed consent, in accordance with human research ethics guidelines. We designed an addition task with seven levels of difficulty, starting from one digit addition (very low) to multi-digit addition (extremely difficult) as shown in Table I.

TABLE I.
TASK DIFFICULTY LEVEL DETAILS.

Task level	Number of digits	Example
Very low (L1)	1&2 digit numbers	45+2
Low (L2)	1&2 digit numbers with 1 carry	54+9
Medium (L3)	2 digit numbers with 1 carry	67+42
Medium-High (L4)	2 digit numbers with 2 carries	39+65
High (L5)	2&3 digit numbers with 1 carry	377+32
Very high (L6)	2&3 digit numbers with 2 carries	76+347
Extremely high (L7)	3 digit numbers with 3 carries	983+748

The EEG signals were recorded from 32 channels mounted in an elastic cap, according to the extended international 10-20 system using an Active Two acquisition system. The experiment was conducted under controlled conditions in an electrically isolated laboratory, with a minimum distance of five meters from power sources to the experiment desk and under natural illumination.The EEG signals were analyzed using three different

non-linear/dynamic measures. They were correlation dimension (CD), Hurst exponent (HE) and approximate entropy (ApEn).

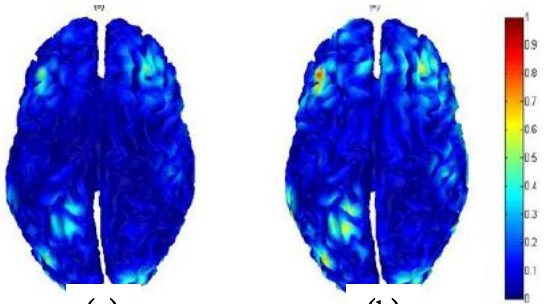


Figure 1. The source maps of two load levels for subject 1; (a) the lowest load (L1), and (b) the most difficult load (L7). Both load levels influence the similar regions more or less but the degree of activation increased as the load level increased.

Results and Discussion: Experimental results show that the values of the measures extracted from the delta frequency band of signals acquired from the frontal and occipital lobes of the brain vary in accordance with the task difficulty level induced (see Figure 1). The values of the correlation dimension increased as the task difficulty increased, showing a rise in complexity of the EEG signals, while the values of the Hurst exponent and approximate entropy decreased as task difficulty increased, indicating more regularity and predictability in the signals.

3.2 Wavelet Complexity Features of EEG Signals for CLM

Experiment: In this study, the use of wavelet-based complexity measures of EEG signals were investigated to evaluate changes in working memory load during the performance of a cognitive task with varying difficulty/load levels. EEG signals were acquired from twelve healthy male subjects; postgraduate students aged between 24-30 years. In the experiment, the participants were asked to do an arithmetic task (an addition task with varying difficulty level, see Table I).

The subjects' EEG signals were recorded using an Active Two system. Each recording contained 32 EEG channels mounted in an elastic cap, according to the extended international 10 - 20 system. A linked earlobe reference was used and impedance was kept under 5 kΩ. The EEG signals were passed through a band-pass filter with cut-off frequencies of 0.1 - 100 Hz and were recorded at an *fs* =256 Hz sampling rate. To select the epochs which contained minimal EMG artifact, each recording was judged by visual inspection. As a result, 70 seconds (out of 90 seconds of each task level recording) for each subject was considered. This portion of the recordings included EOG and ECG artifacts, which were not removed.

Extracted signals were analyzed using wavelet based complexity measures. The wavelet complexity measures associate with four entropic measures: that is Shannon, Tsallis, Escort-Tsallis and Renyi entropies.

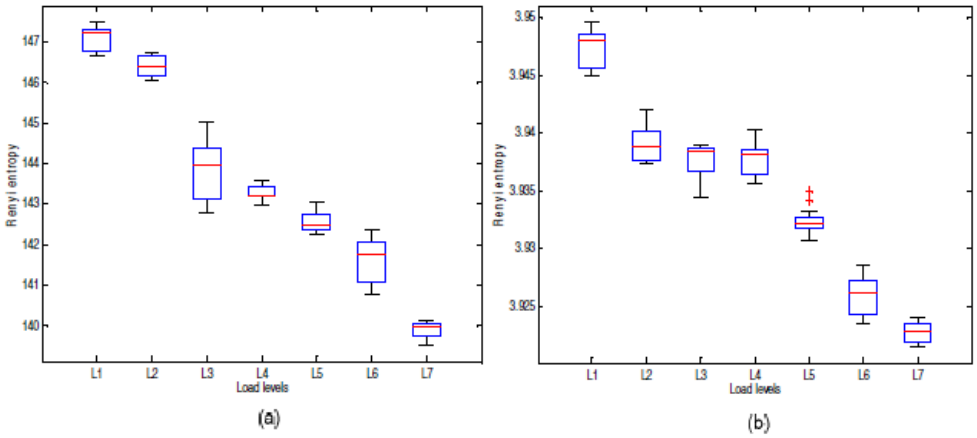


Figure 2. The Renyi entropy variations for (a) q=0.9, (b) q=0.1 with the load levels, for channel F7 of subject 1. On each box, the red mark is the median; the edges of the box are the 25th and the 75th percentiles.

Results and Discussion: As an example, Figure 2 shows the median of the extracted H_{RE} from the frontal channels in scale 5, for channel F7 of subject 1, for two extreme values of q (entropic index); (a) q= 0.9, (b) q= 0.1, in the delta frequency band. As shown, the median of the extracted H_{RE} are able to distinguish the seven task loads better with q closer to 1, as it consistently reveals a decreasing median with increasing task load.

The experimental results demonstrated good discrimination among seven load levels imposed on the working memory with a classification rate of up to 96% using signals recorded from the frontal lobe of the brain. The extracted measures' values show a consistent decrease in the selected channels in two frontal and occipital lobes, as the memory load increases, indicating the EEGs disorder declines while the complexity grows. This illustrates that the brain behaves in a more organized manner characterized by more order and maximal complexity when dealing with higher load levels. The growing complexity can also reflect the higher activation of neural networks involved, as the task load increases.

3.3 Entropy Based Features of EEG Signals for CLM

Experiment: In this study, we investigated the use of entropy-based features (spectral and approximate entropies) of recorded EEG signals to characterize mental load when performing a cognitive task. The participants' EEG signals were recorded using the same method as in the study of Non-Linear Analysis of EEG Signals for CLM and Wavelet Complexity Features of EEG Signals for CLM (six participants were involved in the experiment).

The recorded EEG signals were analysed using following methods: 1) EEG signal source localization using the minimum norm estimate algorithm, 2) sub-band filtering by Discrete Wavelet Transform (DWT), 3) entropy-based feature extraction from the EEG signals.

Results and Discussion: The experimental results demonstrated that the spectral entropy is a good discriminator of mental load level and decreases consistently in accordance with the increased load. The extracted approximate entropy quantifies the irregularity of the EEGs, indicating a decrease in irregularity as the load increases. We also perform EEG source estimation to choose a smaller subset of EEG channels which make the most contribution in the load level discrimination. We conclude that the entropy-based features are capable of measuring the imposed mental load from the selected channels in two brain regions. This may demonstrate that the brain behaves in a more regular or focused manner when dealing with higher task loads. The efficacy of entropy-based features across

frequency subbands was also analyzed in this study.

4. Eye Activity for CLM with Emotion Interference

Eye activity has advantages in CLM. For example, eye activity is more ubiquitous than other modalities; pupillary response and eye blink have been shown to correlate with both visual and aural cognitive tasks; eye activity data collection is less intrusive than other physiological signal data collection. Eye-based CLM is a popular physiological index of cognitive workload that can be used for design and evaluation of adaptive interface in various areas of human-computer interaction (HCI) research.

Eye-based automatic CLM was studied in our research. Three types of eye activity were investigated: pupillary response, blink, and eye movement (fixation and saccade). Eye activity features were investigated in the presence of emotion interference, which is a source of undesirable variability, to determine the susceptibility of CLM systems to other factors.

Experiment: In this study, cognitive load was induced using arithmetic tasks, and the difficulty level was controlled by the number of carries and digits. Emotional interference corresponding to different arousal and valence levels was induced by showing International Affective Picture System (IAPS) images in the task background. The experiment was adapted from those using pupillary response for measuring cognitive load with arithmetic tasks and for measuring arousal with IAPS images.

The participants comprised seven females and eight males, aged 20–48. A total of 82 recordings were obtained from each participant, including 60 samples with both cognitive load and emotion factors, 10 samples with only the cognitive load factor and 12 samples with only the emotion factor. The signal length of each sample was 14 s, during which four task stimuli were systematically presented and time stamped. Figure 3 shows the time line for each task.

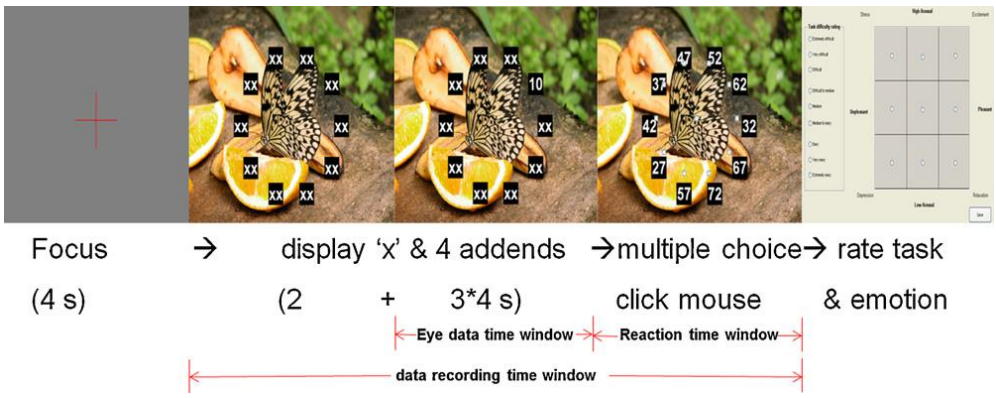


Figure 3. Time line for each task. Each task comprises focusing, image viewing, reading and calculating four addends sequentially, selecting an answer and subjective rating of both task difficulty and emotion.

Results and Discussion: ANOVA test results and multiple regression results revealed important implications for using eye activity for CLM. Pupil size and blink number increased with more difficult tasks, which perfectly matches the literature. Pupil size also increased with higher arousal images regardless of valence, which is also consistent with studies of pupil dilation using visual and auditory stimuli. However, pupil size increased with images of positive valence when a task goal was presented in this study, as the p value was close to 0.05.

The new finding here was that some eye activity feature patterns (notably pupil dilation and blink) for the cognitive load levels were not significantly altered with or without arousal factor in the task-goal driven situation. In contrast, the patterns of features for the arousal level seemed weakened in the pupillary response when cognitive load was induced and there was no arousal effect on the features of blink, fixation and saccade. This result suggests the dominance of cognitive load over emotion in eye features during task performance.

5. Linguistic and Grammatical Features for CLM

Linguistic and grammatical features may be extracted from users' spoken language and analysed for patterns indicating high cognitive load. These features may include speech pauses, self-corrections, repetitions, response latency, and language usage, for example, use of different word categories and parts of speech, such as nouns and pronouns, and grammatical structures. Such features may be collected from users' spoken or written language and are highly unobtrusive. Linguistic features have been regarded as indices of high cognitive load.

Experiment: This research studied 33 members of bushfire management teams working collaboratively in computerized incident control rooms and involved in complex bushfire management tasks. The team members carried out 10 tasks, each about 5 hr in duration, in four states of Australia, including New South Wales, Victoria, Tasmania, and Queensland.

The participants' communication was analyzed for some novel linguistic features as potential indices of cognitive load, which included sentence length, use of agreement and disagreement phrases, and use of personal pronouns, including both singular and plural pronoun types.

Results and Discussion: The experimental results confirmed that while working collaboratively and performing high-cognitive load tasks, people speak more with other team members to manage and share the high task complexity. The results showed that participants, especially those working in a collaborative team environment, consistently use singular pronouns and plural pronouns differently in different task load situations. Specifically, they used significantly more singular pronouns for low-load tasks than for high-load tasks; that is, the lower the cognitive demand, the greater use of singular pronouns. In contrast, they used significantly more plural pronouns for highload tasks than for low-load tasks; that is, the higher the cognitive load, the greater use of plural pronouns. These results support the notion that people actually collaborate and coordinate tasks more with each other during highly complex real-world tasks.

6. Conclusions and Future Work

This research carried out CLM study of three unobtrusive modalities: EEG, eye activity, and linguistic feature based CLM.

In the EEG based CLM, we examined the use of various features (e.g. spectral and approximate entropies, wavelet-based complexity measures, correlation dimension, Hurst exponent) of EEG signals to evaluate changes in working memory load during the performance of a cognitive task with varying difficulty/load levels. Experimental results showed that EEG may be more reliable than self-rating, and capable of distinguishing seven load levels induced under controlled conditions with accuracies exceeding 94%.

In the eye based CLM, three types of eye activity were investigated: pupillary response, blink, and eye movement (fixation and saccade). Results from experiments combining arithmetic-based tasks and affective image stimuli demonstrated that arousal effects were dominated by cognitive load during task execution.

The linguistic feature based CLM was also investigated in this study. Some novel linguistic features were analyzed as potential indices of cognitive load. Results showed that with high load, people spoke more and used longer sentences, used more words that indicated disagreement with other team members, and exhibited increased use of plural personal pronouns and decreased use of singular pronouns.

Future work will include analyzing the cognitive workload based on pupillary response under luminance and emotional changes. Furthermore, understanding the contextual task characteristics and user behavior in interaction can benefit the measurement of cognitive load and development of intelligent systems to aid user task management. The direct and continuous observations of individual tasks via eye activity will be investigated in the future work.

Attachment A

Literature Review on Physiological Measures of Cognitive Workload

Yang Wang, Jianlong Zhou
January 2013
Machine Learning Research Group
NICTA

Contents

1 Introduction

Cognitive (mental) workload is an important issue in various application areas such as human computer interaction, adaptive automation and training, traffic control, performance prediction, driving safety, and military command and control (Byrne and Parasuraman, 1996; Coyne et al., 2009; Grootjen et al., 2007; Wilson and Russel, 2006; Kerick and Allender, 2004). Although numerous approaches have been developed to study cognitive workload or understand how hard the brain is working under various situations, it is still difficult to examine the cognitive workload of a person: "workload is a multidimensional, multifaceted concept that is difficult to define. It is generally agreed that attempts to measure workload relying on a single representative measure are unlikely to be of use" (Gopher and Donchin, 1986). Both theories and models have been proposed to explain cognitive workload. The multiple resource theory models cognitive resource of a person with three different dimensions: perceptual modality, information code, and processing stage (Wickens, 2002). On the other hand, the cognitive load theory models the interaction between limited working memory and relatively unlimited long term memory during the learning process (Sweller, 1988). The theory distinguishes between three types of cognitive workload: intrinsic load, extraneous load, and germane load. The first type is associated with the nature of learning material, while the latter two are influenced by instructional design (Paas et al., 2003).

When a subject or operator is required to perform a given task, cognitive workload could be viewed as the interaction between the demands of the task and the capacity of the subject (Cain, 2007). Such point of view highlights two key issues of mental workload, the subject's capacity and the task demands. The mental workload of a subject tends to increase when the cognitive capacity becomes low, and it tends to increase when the task demands become high. It should be noticed that both subject's capacity and task demands are not necessarily constant values and they may change over time. The capacity of an operator may increase or decrease due to various factors such as training, fatigue, and environment. During a task, an operator can also experience varying levels of workload according to the task difficulty at different stages.

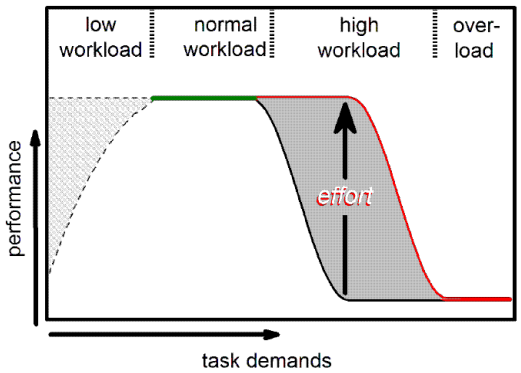


Figure 1. The relationship between task demands, performance, and workload (Veltman and Jansen, 2006).

In recent decades, a great variety of measuring techniques, from simple ones such as questionnaires to complex ones such as functional braining imaging, have been developed to study cognitive workload (Gingell, 2003; Just et al., 2003; Wierwille and Eggemeier, 1993). Generally, these measuring techniques can be divided into three categories: subjective rating, performance measure, and physiological measure (Hart and Staveland, 1988; O'Donnell and Eggemeier, 1986; Wilson et al., 2004). Comparing with subjective rating, the latter two categories provide approaches to assess mental workload in an objective way. One main advantage of objective measurement is that it will not disturb the operation of the subject during the task execution. For performance based measure, the relationship between workload and performance is shown in Figure 1. An operator's performance could be maximized if the task just requires normal mental workload. Meanwhile the performance tends to decline when the task demands become high or even exceed the capacity of the operator. The performance is also influenced by various factors such as attention, expertise, experience, stress, and motivation.

With the advance of modern sensor technologies, more and more physiological measures have been developed for the assessment of cognitive workload. Popular physiological measures used in workload studies include brain wave, eye activity, respiration, heart rate, and speech, etc (Fournier et al., 1999; Scerbo et al., 2001; Yin et al., 2008). Among these techniques, video based workload measures, especially the ones through remote sensing, have attracted increasing attention since they can provide physiological evaluation of cognitive state in a non-intrusive and non-obtrusive way.

Although various studies exhibit the effects of mental workload on physiological measures, no single physiological measure will be sufficient to comprehensively characterize the workload, especially in the case of multidimensional task and/or dynamic circumstances. On the other hand, changes in physiological measures may take place due to a lot of other aspects, such as engagement, fatigue, stress, and environment. Mental workload is just one of these factors influencing physiological measures.

2 Video Based Workload Measures

For the convenience of cognitive workload measurement in different experiments, sensors are selected by the following three usability criteria (Voskamp and Urban, 2009): non-intrusiveness, non-obtrusiveness, and simplicity. Usually a subject does not prefer a device that may invade the human body in any way. Ideally, the applied sensor will not interrupt the operator during the task execution. Moreover, it should not require much effort or training to gather the measurement data.

For the effectiveness of mental workload measures in various cognitive tasks, sensors should also meet the following three technology criteria: sensitivity, efficiency, and compatibility. The selected sensor is required to provide data that is highly correlated to cognitive workload. For online or interactive systems, the collected data needs to

be transferred and processed in real-time. When multiple sensors are applied, the sensors should be easily combined with each other.

2.1 Imaging sensors for workload studies

Based on the sensor selection criteria for cognitive workload study, video camera or imaging sensors have attracted increasing interests during the development of workload measurement techniques. One valuable type of physiological measure involves workload effects on activities of human eye (Sirevaag and Stern, 2000). Especially, video based eye tracker becomes a popular approach for cognitive workload evaluation due to its sensitivity and convenience. Eye tracking data provides important information about human brain activity and autonomic nervous system, and it is highly correlated with subject's mental workload. The visual information is acquired in a non-intrusive (particularly with remote systems, see Figure 2) and continuous way without interfering user's activity during the task performance. Moreover, the video sequences of eye tracking data can be captured with high frame rate (more than 30 frames per second) and processed in real-time. Currently, various eye tracking hardware platforms are commercially available. Bartels and Marshall compared various eye tracking hardware platforms for the measurement of cognitive workload from different manufacturers, such as SR Research, Seeing Machines, SensoMotoric Instruments and Tobii Technology. They found that the pupil recording of each system was precise enough to effectively utilize the Index of Cognitive Activity, one of validated cognitive workload metrics (Bartels and Marshall, 2012).

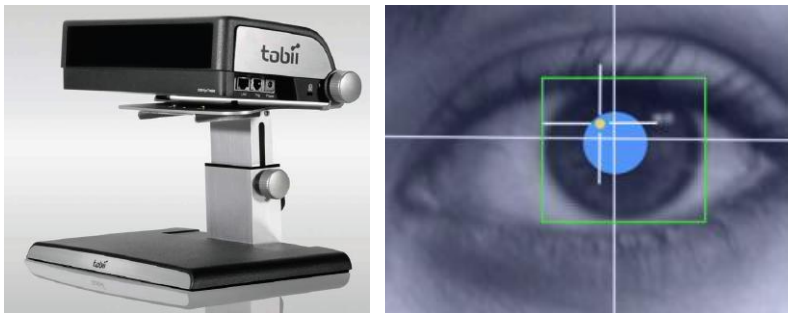


Figure 2. Eye tracker and eye tracking data.

Another type of imaging data, facial skin temperature, has been utilized as a physiological measure in mental workload studies as well. Facial skin temperature shows significant correlation to changes of mental status (Veltman and Vos, 2005). During the task execution, the autonomic nervous system of the subject causes the redistribution of blood flow. Consequently, it will result in the change of local skin temperature. With the use of thermal infrared camera (see Figure 3), the remote sensing of skin temperature can be achieved through measuring the infrared emitted from human body.

On the other hand, advanced brain imaging techniques, such as functional magnetic resonance imaging (fMRI) and near-infrared (NIR) neuroimaging, have also been employed to detect changes in cognitive workload (Callicott et al., 1999; He et al., 2007; Izzetoglu et al., 2005; Sammer 2006). However, due to the constraint of sensing

technology and device, in practice it is hard for those sensors to capture the imaging data in a convenient and non-obtrusive way, which limits their usability as physiological measures of mental workload.

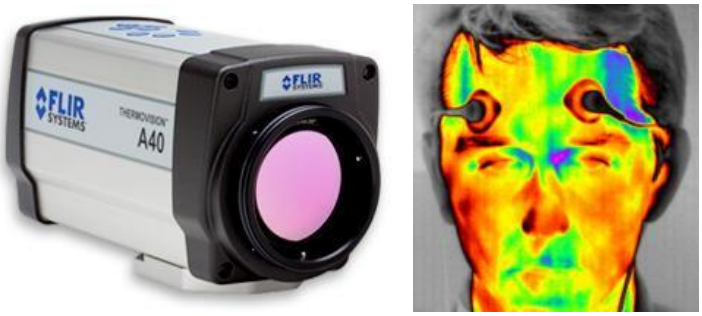


Figure 3. Infrared camera and thermal imaging data.

2.2 Video based measures in cognitive tasks

Eye tracking data provides rich information for cognitive workload assessment. Physiological workload features related to eye activity can be categorized into three classes: eye blink based measures, eye movement (saccade and fixation) based measures, and pupillary response based measures.

In addition to eye activity and facial skin temperature, physical features of human behaviour such as head movement, hand gesture, and facial expression, which can be detected in a non-contact way using video camera, also provide useful information about changes in mental states (Grootjen et al., 2006). Since such physical behaviour measures are relatively less sensitive to cognitive workload, they are usually integrated with other physiological measures to achieve satisfactory performance. Table 1 lists popular video based measures that have been used in cognitive workload studies. Workload research groups working on video based physiological measures include Air Force Research Laboratory, Naval Health Research Center, Human Factors Group of Federal Aviation Administration, TNO Human Factors Research Institute, etc.

To study the effects of mental workload on physiological measures, various cognitive tasks have been designed and tested in both laboratory and real world. The performed tasks include visual, auditory, arithmetic, executive, and complex ones such as driving, traffic control, and flight. Sometimes dual tasks are performed in the workload experiments. It should be noted that multitasking is common for human activities under laboratory and real world environment. For example, even a simple auditory addition task will involve both verbal processing and arithmetic processing; a driving task can be decomposed into at least two subtasks (visual and memory) that require the driver to keep the vehicle on the road and remember the route to the destination. When multiple physiological measures are available, it will be sensible to consider the embedded multimodal information with a composite index for mental workload evaluation (Sciarini and Nicholson, 2009).

Table 1. Video based workload measures

Category	Measure	Explanation
Eye blink based	Blink frequency	The rate of blink times over a time period
	Blink duration	The time interval during the closure of eye
	Blink interval	The time interval between two successive eye blinks
Eye movement based	Fixation number	The times of fixation
	Fixation frequency	The rate of fixation times over a time period
	Fixation duration	The time interval during a fixation
	Saccade rate	The rate of saccade times over a time period
	Saccade extent	The angular distance within a saccade
	Saccade duration	The time interval during a saccade
	Saccade velocity	The angular velocity within a saccade
	Scan path	The trajectory of eye gaze
	Vergence angle	The gaze difference between left eye and right eye
Pupillary response based	Pupil dilation	The increase of pupil size comparing with baseline
	Percentage change in pupil size	The rate of pupil dilation over baseline pupil size
	Index of cognitive activity	Based on changes in pupil dilation (Marshall, 2002)
	Power spectrum	The power spectrum of pupil size data
Skin temperature based	Nose temperature	
	Forehead temperature	
Physical behaviour based	Head movement	
	Facial expression	
	Hand movement	

3 Pupillary response based measures

The correlation between pupillary response and changes in mental workload has been observed for decades (Beatty, 1982). It is known that human eye is regulated by the autonomic nervous system, and pupil diameter will decrease or increase based on autonomic response. Increased pupil diameter is usually observed with an increase in workload demand. Generally, pupil dilation is an important physiological measure of mental efforts and has been widely applied as an effective indicator of cognitive workload.

3.1 Correlation to workload in visual task

Backs and Walrath (1992) evaluated the changes in mental workload when utilizing colour coding for symbolic tactical display in a visual search task. Participants were required to abstract different types of information from the display with varying symbol density. Two pupillary response measures, pupil dilation and constriction-dilation difference, were collected as physiological indices of visual workload. It was found that pupillary response was not only affected by display parameters such as colour coding and symbol density, but also sensitive to the information processing demands of the visual task.

In the experiment of a visuospatial task with varying target density (Van Orden et al., 2001), changes in eye activity based physiological measures were examined during the task. Pupil diameter, together with blink frequency and fixation frequency, were found to be the most relevant eye activity features regarding the target density. Moreover, in the experiment of cognitive task and visual search task (Recarte et al., 2008), the analysis results exhibited that pupil dilation could effectively measure the mental efforts during the cognitive tasks, and it could be used as a physiological predictor of visual impairment as well.

Verney et al. (2001) investigated task-evoked pupillary response in the experiment of a visual backward masking task. The experimental results showed that pupil dilation response became significantly greater during the task condition than during the passive condition of stimulus viewing. Comparing with the non-mask condition, the pupil dilation exhibited significantly increase under the masking condition, especially when the interval between target and mask stimuli was prolonged. As pupillary dilation increased when resource allocation became intensive in the visual task, the experiment demonstrated that the mask could demand extra processing resources when it followed the target by prolonged interval.

Both time domain and frequency domain of physiological data provide useful information for mental workload estimation. The power spectrum of pupillography, especially the band of lower frequency, could be employed as a physiological measure of mental activity as well. Nakayama and Shimizu (2004) studied the frequency information from the task-evoked pupillary response. In the experiment, participants performed visual following task together with/without oral calculation

task under different difficulty levels. Pupil size was recorded as physiological measurement during the task performance. It was found that for the oral calculation task, the power spectrum density of pupil size data increased with higher task difficulty level in the band of 0.1-0.5 Hz and 1.6-3.5Hz, which was consistent with the changes in average pupil size.

Pupil dilation is known to exhibit effects of both the illumination condition of the visual field and the cognitive workload of the person while performing a visual task. Pomplun and Sunkara (2003) investigated effects of cognitive workload and display brightness on pupil dilation and their interaction in the experiment of a gaze-controlled human-computer interaction task. During the visual task, three levels of task difficulty were combined with two levels of background brightness (black and white). The experimental results showed that under both black and white background conditions, the pupil area exhibited significant increase when workload demands became higher. However, under bright background even the pupil area corresponding to high level of task difficulty was significantly smaller than the pupil area corresponding to low level of difficulty under black background. Hence comparing with the task difficulty, the background brightness actually resulted in greater variation of pupil area.

Klingner et al. (2011) investigated the effect of aural vs. visual task presentation on pupil dilation for cognitive load. Three tasks spanning from mental multiplication, digit sequence recall to vigilance were performed in the study. It was found that the patterns of pupil dilation were similar for both aural and visual presentation for all three tasks, but the magnitudes of pupil response were greater for aural presentation. Accuracy was higher for visual presentation for mental arithmetic and digit recall. The findings suggest that cognitive load is lower for visual than for aural presentation.

3.2 Correlation to workload in driving task

Marshall (2002) proposed a physiological measure of cognitive workload, index of cognitive activity (ICA), from changes in pupil dilation. ICA would measure abrupt discontinuities in pupil diameter and try to separate pupil's reflex reaction to changes in light from the reflex reaction to changes in workload. In the cognitive workload study with a simulated driving task (Schwalm et al., 2008), the experimental results showed that ICA increased when workload demands became high, which was induced by performing lane change manoeuvre or additional secondary task. The study exhibited the feasibility of ICA as a physiological measure of mental workload while driving.

In a dual task experiment, Tsai et al. (2007) examined pupillary response when subjects performed driving task and auditory addition task simultaneously. It was found that pupil dilation was significantly greater when subjects were performing well in the auditory task than when subjects were performing poorly.

In another experiment of dual task, Palinko et al. (2010) also studied the pupillary response with remote eye tracker. The subjects performed simulated vehicle driving

as well as spoken dialogues. In the experiment, pupil size data acquired from remote eye tracker was used for the evaluation of the driver's cognitive load. During the task, the physiological measure based on pupillary response exhibited significant correlation to those measures based on driving performance. A pupillary response based measure of cognitive load, mean pupil diameter change rate, was proposed to analyse workload changes with small time scales. The experimental results demonstrated the reliability of physiological measures obtained through remote eye tracking for cognitive load estimation.

3.3 Correlation to workload in arithmetic/memory task

Murata and Iwase (1998) assessed mental workload based on the fluctuation of pupil area. In the experiment, a mental division task and a Sternberg memory search task were carried out with the controlling of respiration. During the task, the number of digits and the size of memory set were used to manipulate the mental workload level induced by task demands. For each subject, the autoregressive power spectrum of pupil area was used for cognitive workload assessment. It was found that the ratio of power at low frequency band (0.05-0.15Hz) over power at high frequency band (0.35-0.4Hz) increased with higher level of task difficulty for both the arithmetic task and the memory task. The experimental results indicated that the fluctuation rhythm of the pupil area could be used as an effective physiological index to evaluate mental workload.

Klingner et al. (2008) examined the pupil measuring capability of video based eye tracker for cognitive workload evaluation. In the experiment of several tasks including arithmetic and memory ones, subtle changes of pupil size in the task-evoked pupillary response were detected using remote eye tracker. Comparing with the results in earlier studies, it was found that cognitive workload could be effectively measured through remote eye tracking. Moreover, the experimental results exhibited the feasibility of analysing the timing and magnitude of short-term pupillary response based on the collected eye tracking data, which could provide more details about changes in cognitive workload.

Xu et al. (Xu et al., 2011a; Xu et al., 2011b) studied the characteristics of pupillary responses at different stages of cognitive process when performing arithmetic tasks under luminance changes. The arithmetic tasks in the study have 4 levels of difficulty, and each level of task difficulty is combined with 4 levels of background brightness, which results in 16 different trial types in total. The results showed that a small pupil diameter is usually observed under brighter background, and the pupil diameter often increases when the task difficulty level becomes high for each background brightness level. The further fine-grained analysis for the experimental results showed that the measurement values of the pupil diameter increase as the task difficulty increases under the influence of luminance changes.

3.4 Correlation to workload in other tasks

In an early study, Beatty (1982) investigated task-evoked pupillary response in the experiments of various tasks such as language processing, reasoning, and perception. Pupil dilation was exhibited as a reliable physiological measure of mental state or processing load during the task performance. Similarly, in the recent experiment of a combat management task involving target identification (Greef et al., 2009), pupil dilation also increased when cognitive workload became high.

In the experiment of air traffic controller task (Ahlstrom and Friedman-Berg, 2006), mean pupil diameter was employed as the physiological measure of mental workload. It was found that comparing to when using a dynamic forecast tool, the mean pupil diameter became significantly larger when using a static forecast tool. The experimental results indicated that the use of static tool led to higher cognitive workload. In another experiment of a video game task (Lin and Imamiya, 2006), it was also found that pupil size increased when task difficulty changes from low level to high level.

In a study of anaesthetists' workload fluctuations during full-scale simulator sessions (Schulz et al., 2011), pupil diameter was used as one of the physiological measures of workload. It is found that pupil diameter and heart rate increased simultaneously as the severity of the simulated critical incident increased.

For interruption management in interactive systems, notifications delivered during the period of lower mental workload would become less interruptive (Iqbal et al., 2004). Bailey and Iqbal (2008) empirically examined changes in mental workload during goal-directed interactive tasks including reading comprehension, mathematical reasoning, product searching, and object manipulation. Percentage change in pupil size was used as the task-evoked pupillary response for continuous workload measurement. The experimental results showed that workload would decrease at subtask boundaries, and the decrement would be greater at boundaries when the operators accomplished large chunks of the interactive task. For operators of interactive systems, pupillary response was exhibited to be a meaningful index of mental workload during the execution of a hierarchical task.

Although mental workload has been exhibited to decrease at subtask boundaries, it has not been examined for subtasks requiring different devices such as notebook computer and mobile phone. Tungare and Perez-Quinones (2009) proposed to study the changes in mental workload for multi-device personal information management. In an ongoing experiment, participants would perform information collection tasks using different devices. Pupil diameter would be monitored to provide continuous measurement of workload.

Existing software analysis tools usually can generate the graph of pupillary response over time and playback the video of user's screen interaction, but may not allow the response data to be interactively explored with regard to the task execution model. To facilitate analysis of pupillary response data in relation to the hierarchical structure of the task, Bailey et al. (2007) developed an interactive analysis tool to analyse mental workload if the task could be decomposed into hierarchical subtasks. The workload

data was precisely aligned to the corresponding task execution model during the analysis.

4 Eye blink based measures

Pervious research work has exhibited that eye blink is a useful measure of mental workload (Fogarty and Stern, 1989), especially for workload demands associated with visual tasks. In several experiments using either electro-oculogram (EOG) or video eye tracker, blink rate decreases with an increase in cognitive workload; increase of blink interval is observed with increased mental workload; meanwhile blink duration tends to decrease against more intense processing load. Such blink based physiological response help human eye to save more time to handle visual information during the task performance.

4.1 Correlation to workload in visual task

Van Orden et al. (2001) investigated changes in various eye activity based measures in a visuospatial memory task with varying target density. Two eye blink based measures, blink frequency and blink duration were monitored during the task. In the experiment, subjects were required to recognize and remember each target's identification (friend or enemy) on the display for appropriate action (fire or not) when the targets were approaching. It was demonstrated that both blink frequency and blink duration declined with increasing target density during the visuospatial memory task.

Recarte et al. (2008) examined the concurrent validity of eye activity based physiological measures for mental workload evaluation. The participants performed single cognitive task and dual task (cognitive task and visual search) in the experiment. Under single task condition, blink rate and pupil dilation showed concurrent validity for mental workload assessment. However, the blink rate exhibited opposite effects under the dual task condition. The blink rate increased when the mental workload of cognitive task became high, meanwhile the blink rate deceased when visual demands became high.

Startle eye blink reflex is also affected by workload demands during visual task. Neumann (2002) studied changes in startle blink during a continuous visual task with different levels of mental workload. In the experiment, subjects performed a single task of visual horizontal tracking or a dual task of both visual horizontal tracking and visual gauge monitoring. The startle blink reflex was evoked by a noise burst during the task execution. Experimental results exhibited that compared with pre-task and post-task conditions, startle blink was suppressed during the task performance. Moreover, compared with the single task condition, the suppression became more significant under the dual task condition. The startle blink rate and other measures such as subjective rating and heart period showed concurrent validity for different workload levels, which indicated that startle blink could be a useful physiological measure of mental workload during the visual task.

4.2 Correlation to workload in flight task

Veltman and Gaillard (1998) investigated the sensitivity of various physiological indices, including eye blinks, in simulated flight tasks. In the experiment, subjects simultaneously performed a continuous memory task during the flight. Eye blink based measures including blink interval, blink duration, closing time and amplitude were monitored during the experiment. Comparing with the measurement data during rest status, blink interval increased and blink duration decreased when subjects performed flight tasks. In addition, blink interval increased and blink duration decreased when subjects were processing more visual information during the flight. On the other hand, the experimental results also showed that the blink interval decreased with increasing difficulty level of the memory task. The decrement was probably due to sub-vocal activity that stimulated the muscles of eyelid and resulted in increased eye blinks.

Similar results were found by Wilson (2002) in the experiment of real flight task. For each pilot, eye blink was recorded as one physiological measure during a flight with both visual rule and instrument rule conditions. The results showed that blink rate decreased when the segments of flight became more visually demanding. In the experiment, each pilot repeated the same task to examine the reliability of the physiological measures, and similar response data was obtained for the two rounds.

4.3 Correlation to workload in traffic control task

Brookings et al. (1996) examined the sensitivity of physiological response to changes in cognitive workload during simulated air traffic control task. In the experiment, eye blink rate exhibited significant effects of task difficulty. The level of task difficulty was manipulated by varying traffic volume and traffic complexity. Eye activity based physiological measures including blink rate were monitored during the traffic control task. The experimental results showed that blink rate decreased with increasing cognitive load.

Ahlstrom and Friedman-Berg (2006) investigated the effect on cognitive workload with/without the use of weather display during air traffic controller task. In the experiment, blink frequency and blink duration were used as two of the physiological workload measures. It was found blink duration became significantly shorter when controllers operated without using weather display, corresponding to a higher level of controller workload. The experimental results also indicated that comparing with subject rating, eye activity based features was relatively sensitive to the variation of mental workload at system interaction stages.

4.4 Correlation to workload in other tasks

In an experiment of dual task, Tsai et al. (2007) investigated changes in eye activities while subjects performed driving task and paced auditory serial addition task. In the experiment, two eye blink based physiological measures, blink frequency and blink duration were recorded. Experimental results exhibited that comparing with the

measurement data in the single task of driving, blink frequency increased in the dual task of both driving and auditory addition. In another experiment of complex decision making task (Boehm-Davis et al., 2006), the results exhibited that eye blinks would be suppressed during cognitive processing comparing to when the processing was accomplished.

Ryu and Myung (2005) employed multiple physiological measures to evaluate the mental workload in a dual task with different difficulty levels. In the experiment, the subjects simultaneously performed a tracking task of simulated instrument landing and mental arithmetic task of adding pairs of numbers. Eye blink interval was employed as one physiological measure for mental effort assessment in both tasks. It was found that the blink interval revealed sensitivity to the changes in mental workload for the tracking task, but not for the arithmetic task.

Zheng et al. (2012) utilized a paper assessment instrument (National Aeronautics and Space Administration Task Load Index, NASA TLX) to evaluate surgeons' mental workload through examining their eye blinks. Surgeons' eye blinks were video-recorded using a head-mounted eye-tracker while the surgeons performed a laparoscopic procedure on a virtual reality trainer. It shows that surgeons who blinked infrequently reported a higher level of frustration (46 vs. 34, $P = 0.047$) and higher overall level of workload (57 vs. 47, $P = 0.045$) than those who blinked more frequently.

5 Eye movement based measures

Eye movement mainly consists of two forms of activity: fixation and saccade. During the visual scan, human eyes are directed to interesting areas where fixations occur. A fixation is a steady focus of the eye, inputting detailed information of the visual stimulus into human vision system. The movement from one fixation stimulus to another is defined as a saccade. Previous studies revealed correlations between changes in mental workload and properties of eye movement (May et al., 1990). For example, the increase in fixation time has been observed with the increase of mental workload. In several experiments saccade based measures such as saccade speed also exhibited sensitivity to changes in mental workload.

5.1 Correlation to workload in visual task

In the task of visual search of symbolic displays (Backs and Walrath, 1992), number of eye fixation, fixation duration, and fixation frequency were employed as eye movement based physiological indices. It was found that the number of eye fixations was affected by both colour coding and symbol density. In the experiment participants made fewer fixations to search colour-coded displays than monochrome displays, and fewer fixations to search low-density displays than high-density displays. Moreover, compared to when searching monochrome displays, fixation duration became shorter and fixation frequency became higher when searching colour-coded displays.

In the visuospatial memory task of target identification (Van Orden et al., 2001), the task difficulty was manipulated by varying the number of targets presented on the display. Physiological measures including fixation frequency, dwell time, and saccade extent were recorded for each participant in the experiment. It was found through nonlinear regression analysis that among the eye movement based measures, fixation frequency revealed significant correlativity to the target density in the visuospatial task.

Frequency information of eye movement also provides a useful physiological index of mental workload. Nakayama and Schimizu (2004) performed frequency analysis of eye movement data in both single task of ocular following and dual task of ocular following and oral calculation. After correcting the artefacts of eye blinks in saccadic eye movement, cross spectrum density, which exhibits relationship between horizontal and vertical eye movement, was employed as a workload measure. Given the eye movement data of different task difficulty levels, the cross spectrum density exhibited significant differences between them in the frequency band of 0.6-1.5Hz.

5.2 Correlation to workload in driving/riding task

In the experiment with a dual task of driving and auditory addition (Tsai et al., 2007), three physiological measures of eye movement, including fixation frequency, fixation duration, and horizontal vergence, were assessed as the indicator of cognitive workload. Comparing to when the subjects performed poorly in the auditory task, the horizontal vergence increased when subjects performed well. Although there was no significant change in fixation frequency, it was found that fixation duration before incorrect responses of auditory addition were significantly shorter than fixation duration before correct responses in the dual task. The experimental results indicated that eye movement based measures could be utilized to both evaluate cognitive load and predict task performance in real-time.

In the experiment of motorbike riding task, Di Stasi et al. (2009) studied the relationship between cognitive workload and risk behaviour. Eye movement based measures including saccadic number, saccadic amplitude, saccadic duration, peak saccadic velocity, fixation number, fixation duration were used as physiological indices of mental workload. The experimental results showed that comparing with low-risky participants, the cognitive workload became higher for high-risky participants, meanwhile the peak saccadic velocity could be used as a reliable physiological index of risk behaviour.

5.3 Correlation to workload in traffic control task

In an experiment of air traffic control task (Brookings et al., 1996), subjects performed simulated traffic control tasks with varying traffic volume and traffic complexity. Two eye movement based workload measures, saccade rate and amplitude, were recorded together with other physiological measures during the control task. However, the saccade measures did not demonstrate significant effects of task difficulty or traffic complexity in the experiment.

Di Stasi et al. (2010) studied the effects of mental workload on eye movement based indices in simulated air traffic control task. In the experiment, participants performed multitasks with three levels of task difficulty according to the cognitive resource requirement. Three eye movement based physiological measures, saccadic amplitude, saccadic duration, and saccadic peak velocity, were recorded using video eye tracker. Experimental results showed that the peak velocity decreased with increasing task difficulty, indicating the sensitivity of saccadic movement to changes in mental workload.

5.4 Correlation to workload in other tasks

Lin and Imamiya (2006) explored the multimodal information of workload measures for usability evaluation. Multiple physiological measures, including fixation number, fixation duration, scan path length, are recorded to estimate cognitive workload when subjects were performing a video based action-puzzle game task. In the experiment, eye movement data exhibited correlation to mental workload level. It was found that mean values of three eye movement based workload measures increased when the task difficulty changed from low level to high level. Saccade speed also exhibited correlation with heart rate variability during the game task. Moreover, a composite physiological measure combining eye fixations with hand movement (mouse clicks) was proposed to improve the evaluation of task performance.

In the experiment of a combat management task requiring target identification and weapon engagement, Greef et al. (2009) investigated three aspects of eye movement, fixation time, saccade distance, and saccade speed, for objective assessment of mental workload. To examine their correlativity with changes in workload, these features of eye activity were monitored by video eye tracker under different levels of mental workload. The experiment results exhibited that fixation time significantly increased when the mental workload became high. Meanwhile saccade distance and saccade speed did not exhibit any significant effects.

6 Skin temperature based measures

Facial skin temperature can be employed as a type of non-intrusive, non-obtrusive, and real-time physiological measure for mental workload assessment. It has received increasing attention in cognitive workload studies as the cost of thermal infrared camera decreased in recent years. Especially, the skin temperature drop of nose area with increased mental workload has been observed in a few studies.

Veltman and Vos (2005) examined the variation of subject’s facial skin temperature in a continuous memory task with two difficulty levels. The experimental results demonstrated the correlation between nose skin temperature and changes in mental workload. To enhance the sensitivity and accuracy, the facial skin temperature could be integrated with other physiological measures for cognitive workload evaluation.

Or and Duffy (2007) also studied changes in facial skin temperature for automated mental workload assessment. In the experiment, subjects performed driving test under different traffic conditions (city/highway) in simulator or real vehicle. Mental arithmetic test was used as a secondary task. Both forehead temperature and nose temperature were monitored during the experiment. It was found that under all simulator test conditions, nose skin temperature dropped significantly after the driving. The dual task of driving and arithmetic resulted in a greater nose temperature drop than the driving only task. In addition, the experimental results exhibited a significant correlativity between the nose skin temperature and the subjective rating of mental workload. Comparing with the real driving task, the simulated driving task had a higher subjective rating and it was observed with a greater change of nose skin temperature.

Previous research work on facial skin temperature has revealed its correlation to the variation of mental workload. However, it has also been noticed that the skin temperature based measures may not achieve sufficient sensitivity, especially for complex tasks or practical applications. Consequently, the combination of skin temperature and various other measures has been proposed to improve the performance of workload assessment. Wang et al. (2007) presented a composite workload index using three video based physiological measures, facial skin temperature, eye blinks, and pupil dilation. All the measures could be unobtrusively captured in real-time for workload evaluation.

7 Linguistic feature based measures

Besides the easy understandable behavior measures for cognitive load such as eye blinking and movement, linguistic and grammatical features may also be extracted from users' spoken language and analyzed for patterns indicating high cognitive load. These features may include speech pauses, self-corrections, repetitions, response latency, and language usage. Such features can be collected from users' spoken or written language and are highly unobtrusive (Khawaja et al., 2012).

Various linguistic features are examined as indices of high cognitive load. Some researches (Berthold and Jameson, 1999; Mueller et al., 2001; Jameson et al., 2006: Khawaja et al., 2008) showed that some speech features are related to a person's cognitive load levels, such as filled pauses and the number of sentence fragments, and tried to recognize cognitive load levels from a number of high level features by using Bayesian network (Mueller et al., 2001; Jameson et al., 2006). Word frequency and use of first-person plurals (Sexton and Helmreich, 2000) are also used to estimate cognitive load.

Demberg and Sayeed (2011) used linguistic cognitive load in a speech-driven user interfaces for automotive drivers. In this work, measures of language complexity for cognitive load are used to modulate the complexity of driver's user interface with ongoing driving conditions.

Khawaja et al. (Khawaja et al., 2009; Khawaja et al., 2012) investigated linguistic features for measurement of cognitive load in complex bushfire management tasks. In these tasks, bushfire management teams working collaboratively in computerized incident control rooms. The participants' communication was analyzed for linguistic features as potential indices of cognitive load, which included sentence length, use of agreement and disagreement phrases, and use of personal pronouns, including both singular and plural pronoun types. The study showed that with high cognitive load, people spoke more and used longer sentences, used more words that indicated disagreement with other team members, and exhibited increased use of plural personal pronouns and decreased use of singular pronouns.

Linguistic feature based measures are usually speaker-dependent and need manually labelled data, which limits the automatic measurement of cognitive workload.

8 Speech signal based measures

Similar to video based cognitive workload measures, speech spectral/temporal patterns can also be employed as a type of non-intrusive, non-obtrusive, and real-time physiological measure for mental workload assessment.

Yin et al. (2008) investigated speaker-independent approaches by utilizing speech signal process and classification techniques. Speech features such as Mel-Frequency Cepstral Coefficients (MFCC) and temporal information are employed to automatically monitor a person's cognitive workload. To capture the temporal information of speech features, three different approaches are used: Delta cepstrum, Acceleration and shifted Delta Cepstra. The investigation achieved 71.1% and 77.5% accuracy on two different tasks.

As MFCCs (Yin et al., 2008) do not prove with any insight into how cognitive load affects the speech spectrum, glottal features were then investigated to link cognitive load to the speck production system (Yap et al., 2010; Le et al., 2010).

Yap et al. (2011a) employed acoustic voice source features extracted from the speech spectrum (or cepstrum) for cognitive load classification. Pre and post-processing techniques were used to improve the estimation of the cepstral peak prominence (CPP). The results showed that CPP is a promising cognitive load classification feature that outperforms glottal flow features.

Yap et al. (2011b) further employed speech formant frequency-based features for automatic cognitive load classification. The investigation found that the slope, dispersion, and duration of vowel formant trajectories exhibit changes under different cognitive load conditions, and therefore are used in vowel-based classification for cognitive load measurement. The results show that the performace of frame-based formant features in 2-class and 3-class utterance-based cognitive workload classification is comparable with that of baseline MFCC features. Yap et al. (2011a) also used score-level fusion of the CPP-based classification with the formant

frequency-based system and yielded a final improved accuracy of 62.7% in cognitive load classification.

Le et al. (2011) investigated the use of speech's spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features for automatic cognitive load measurement. They found that the spectral centroid features consistently and significantly outperform a baseline system employing MFCC, pitch, and intensity features. They reported that the fusion of an SCF based system with an SCA based system results in a relative reduction in error rate of 39% and 29% for two different cognitive load databases.

9 Electroencephalogram (EEG) based measures

Electroencephalography (EEG) is a noninvasive neuroimaging technique widely used for measuring cognitive workload. EEG contains useful information about various physiological states of the brain and can be very efficient for understanding the complex dynamical behavior of the brain, if interpreted correctly (Hasan, 2009). It can provide continuous and on-line assessment of cognitive load at all levels.

Various methods are used to extract different features of EEG to measure and classify cognitive load. This includes spectral features such as power spectral density (PSD) or the averaged power, maximum/log power spectra, sub-band entropy, and autoregressive model (Diez et al., 2008; Zarjam et al. 2011).

Zarjam et al. (2011) further investigated other spectral features of EEG signals for the measurement of cognitive load in a reading task. The study showed that a set of spectral features − the spectral entropy, weighted mean frequency and its bandwidth, and spectral edge frequency − is able to discriminate the cognitive load levels effectively. The study also found that combination of various features into the classification for cognitive load resulted in better performance compared to one feature taken alone.

Anderson et al. (2011) measured cognitive load based on EEG signals across multiple visualization types. Spectral characteristics of EEG signals such as the alpha (7.5 – 12.5 Hz) and theta (4 – 7.5 Hz) frequency bands are used to reflect cognitive and memory performance. It is indicated that the Box plot and the Density plot used in the study incurred the lowest cognitive load scores, while the Violin and Interquartile plots induced the highest cognitive load. In the study, the Violin and Interquartile plots have greater visual complexity than the Box plot and Density plot.

In an arithmetic task, Zarjam et al. (2012a) employed epectral and approximate entropied of EEG signals to characterize cognitive load. It is demonstrated that the spectral entropy decreases consistently in accordance with the increased load. The extracted approximate entropy quantifies the irregularity of the EEGs, indicating a decrease in irregularity as the load increases. In an another arithmetic task, Zarjam et al. (2012b) used three different non-linear/dynamic measures of correlation dimension, Hurst exponent and approximate entropy of EEG signals to measure cognitive load.

The study showed that values of the correlation dimension increase as the task difficulty increases, while the values of the Hurst exponent and approximate entropy decrease as task difficulty increases.

10 Galvanic Skin Response (GSR) based measures

Galvanic Skin Response (GSR) has recently attracted researchers' attention as a prospective physiological indicator of cognitive load. It is also referred to as electrodermal activity (EDA). GSR is a measure of conductivity of human skin, and can provide an indication of changes in human sympathetic nervous system. Similar to EEG, GSR is also a noninvasive technique for the measurement of cognitive load.

Shi et al. (2007) evaluated users' stress and arousal levels while using unimodal and multimodal versions of the same interface in a traffic control management study. The results showed that mean GSR significantly increase when task cognitive load level increases. Moreover, users' GSR readings are found to be lower when using a multimodal interface, instead of a unimodal interface.

Haapalainen et al. (2010) assessed mean, variance and median of GSR against two cognitive load levels. They did not obtain any satisfactory results for GSR and explained that it might be related to the tasks type or their GSR sensors might not have been sensitive enough.

Son and Park (2011) estimated driver's cognitive load using driving performance and skin conductance level as well as other measures in a driving simulator. The results showed that the skin conductance level provides clear changes associated with difficult level of cognitive workload. It was able to identify driver's cognitive load complexity with high accuracy.

In designed arithmetic and reading tasks, Nourbakhsh et al. (2012) examined temporal and spectral features of GSR against different task difficulty levels. The results show the strong significance of the explored features, especially the spectral ones, in cognitive workload measurement in the two studied experiments.

11 Pen input feature based measures

Writing activities usually require the focused attention of writers, and features of this experience can help to understand cognitive load of writers. Ruiz et al. (2007) examined changes in trajectory velocity and shape-degradation of pen-gesture features as possible indices of cognitive load. They found possible trends in gesture kinematics occurred when switching to high cognitive load in tasks where cognitive load increases continuously. They also observed trends of increased degeneration of gesture shapes as cognitive load increases.

Yu et al. (2011) analysed cognitive load by using the orientation of the pen and the pressure of the pen-tip from digital writing samples. Gaussian Mixture Models (GMM) were used to classify cognitive load levels from handwriting data sets. The results

showed that the pen orientation and pressure reflected cognitive load variation well, and the significant improvement in cognitive load classification from 52.8% to 75.4% validated the effectiveness of sample selection using altitude.

12 Noisy factors in workload measures

Although a number of studies exhibited empirical evidence that eye activity based physiological measures could be used as an effective indicator of mental workload increase, the measures may fail to evaluate workload under complicated situations. For example, pupil dilation could be influenced by experimental environment like illumination condition. In addition to the experiment involving background lightness (Pomplun and Sunkara, 2003), Kramer (1991) reported the failure of workload measure due to factors unrelated to the cognitive task, such as changes in ambient illumination or screen luminance, which might give rise to greater variation of pupil size. Ganguly (2012) studied the pupillary response to variation in both cognitive workload and luminance. The results suggested that task difficulty had a stronger effect on pupil size for the bright background than for the dark background. In an experiment on the effects of perceptual/central and physical demands on physiological measures (Backs et al., 1994), it was found that physiological measures would be more sensitive to physical demands than to perceptual/central demands. In another experiment study of Sternberg memory search task (Van Gervan et al., 2003), the analysis results also demonstrated effects of aging on pupillary response. Moreover, to evaluate the usability of eye tracking data for cognitive workload measurement, Pomplun and Sunkara (2003) studied the distortion of pupil size caused by eye movements. The pupil size observed by the eye tracking camera would be affected by the gaze angle of the user. The eye tracking system was calibrated based on neural network to correct the geometry distortion of pupillary response data.

Zekveld et al. (2011) evaluated the influence of age, hearing loss, and cognitive ability on the cognitive load during listening to speech presented in noise. Cognitive load was assessed by examination of pupil dilation. The results showed that the pupil response systematically increased with decreasing speech intelligibility. Ageing and hearing loss were related to less release from effort when increasing the intelligibility of speech in noise. In difficult listening conditions, these factors may induce cognitive overload relatively early or they may be associated with relatively shallow speech processing. Better text reception thresholds and larger word vocabulary were related to higher mental processing load across speech intelligibility levels.

Video based physiological measures can also be influenced by a variety of affective factors including anxiety, engagement, fatigue, and stress (Chen, 2006; Pavlidis et al., 2000; Prinzel et al., 1999). For example, eye blinks, heart rate variability, or electroencephalogram (EEG) could be used to evaluate engagement and fatigue as well (Heishman and Duric, 2007; Zhang et al., 2008). Genno et al. (1997) investigated the changes in facial skin temperature caused by subject's stress or fatigue during the task. In the experiment of a task inducing stress, the nose skin temperature exhibited significant drop when the task started or an unexpected emergency alarm took place. Moreover, the nose skin temperature dropped significantly as well in the experiment

of another task inducing fatigue. Meanwhile, Puri et al. (2005) also exhibited the correlation between forehead temperature and emotional state through thermal imaging.

Although it would be ideal to find a general model of human cognitive workload, mental workload could be personal characteristics of each subject. Thomas et al. (2009) studied personalized mental workload for exercise intensity measure. In the experiment, ratio of non-blink to blink frames and pupil radius were detected for each participant during different exercise tasks. It was suggested that due to non-stationary and nonzero-state nature of human being system, mental workload should be modelled individually and adaptively.

Table 2. Cognitive task–physiological measure matrix

Task	BF	BI	BD	FF	FD	SD	SS	PD	PC	PS	IC	ST	HM
Air traffic control task (Ahlstrom and Friedman-Berg, 2006)	●		●			●		●					
Air traffic control task (Brookings et al., 1996)	●					●							
Air traffic control task (Di Stasi et al., 2010)						●	●						
Auditory two-back task (Guhe et al., 2005)	●							●					●
Cart driving and stationary bike exercise (Thomas et al., 2009)								●					
Cognitive task and visual search task (Recarte et al., 2008)	●							●					
Combat management task (Greef et al., 2009)					●	●	●	●					
Continuous memory task (Veltman and Vos, 2005)												●	
Division task and Sternberg memory search (Murata and Iwase, 1998)										●			
Driving task and auditory addition task (Tsai et al., 2007)	●		●	●	●			●					
Driving task and secondary task (Schwalm et al., 2008)											●		
Driving task and spoken task (Palinko et al., 2010)								●					
Driving task and verbal/spatial-imagery task (Zhang et al., 2004)								●					
Document editing, email classification, route planning (Bailey and Iqbal, 2008)									●				
Flight task and memory task (Veltman and Gaillard, 1998)		●	●										
Flight task with visual/instrument flight rule (Wilson, 2002)	●												
Gaze-controlled interaction task (Pomplun and Sunkara, 2003)								●					
Language, visuospatial, and executive processing (Just et al., 2003)								●					
Mental arithmetic, short-term memory, aural vigilance (Klingner et al., 2008)								●					
Motorbike riding task (Di Stasi et al., 2009)				●	●	●	●	●					
Ocular following and oral calculation (Nakayama and Shimizu, 2004)								●		●			

Table 2. Cognitive task–physiological measure matrix (continued)

Task	BF	BI	BD	FF	FD	SD	SS	PD	PC	PS	IC	ST	HM
Reading, reasoning, searching, and object manipulation (Iqbal et al., 2004)									●				
Simulated/real driving task and mental arithmetic task (Or and Duffy, 2007)												●	
Tracking task and mental arithmetic task (Ryu and Myung, 2005)		●											
Tracking task and mental arithmetic task (Wang et al., 2007)		●						●				●	
Video game (action-puzzle) task (Lin and Imamiya, 2006)				●	●			●					●
Visual backward masking task (Verney et al., 2001)								●					
Visual horizontal tracking and visual gauge monitoring (Neumann, 2002)	●	●											
Visual search of symbolic displays (Backs and Walrath, 1992)				●	●			●					
Visuospatial memory task (Van Orden et al., 2001)	●		●	●	●	●		●					

Physiological measures. BF: blink frequency, BI: blink interval/latency, BD: blink duration, FF: fixation frequency, FD: fixation duration, SD: saccade distance/extent, SS: saccade speed, PD: pupil diameter/dilation, PC: percentage change in pupil size, PS: power spectrum, IC: index of cognitive activity, ST: skin temperature, HM: head/hand movement.

13 Multimodal measures and data fusion

Although physiological measures have exhibited reliable sensitivity to the variation of mental efforts when operators experience different levels of task demands, it is generally agreed that no single physiological measure can comprehensively describe cognitive workload. For example, in an experiment of actual flight scenario (Hankins and Wilson, 1998), eye activity only showed sensitivity to workload during flight segments that were visually demanding, meanwhile heart rate and EEG respectively showed sensitivity during flight segments of instrument rule and those requiring mental calculation. The experimental results demonstrated the multiple physiological measures could provide unique and non-overlapping information about subject's mental workload.

As multitasking is common in human activities, different subtasks may have different effects on individual physiological measures. In terms of the multiple resource theory for cognitive workload, the processing resource indexed by one video based physiological measure could be different from those indexed by other types of physiological measures. Table 2 lists recent research work using physiological measures for workload evaluation in various cognitive tasks. Multiple workload measures, especially physiological measures, could provide a comprehensive picture

of the processing demands during the task execution. To further increase the performance of cognitive workload assessment, it is reasonable to combine different video based measures and/or other physiological measures.

13.1 Multiple measures vs. single measure

The sensitivity of individual physiological measures to workload demands could be very different. For example, in the experiment of different flight tasks (Veltman and Gaillard, 1998), cognitive workload measures including heart period, blood pressure, respiration, and eye blinks were recorded during the task. Although all the measures showed the difference between rest and fight, only heart period was sensitive to all the difficulty levels in the tunnel fight task.

Lin and Imamiya (2006) studied composite physiological measure through integrating eye movement and hand movement for mental effort evaluation when subjects performed a video game task. Although single physiological measures could only distinguish between the low difficulty level and high difficulty level, the composite measure was able to detect the variation of mental efforts for all the difficulty levels of the game task.

Similarly, Ryu and Myung (2005) showed that in the experiment with a dual task of tracking and arithmetic, none of the three physiological measures, including alpha suppression of brain activity, eye blink interval, and heart rate variability was able to identify the variation of the mental workload for both tasks. The alpha suppression was sensitive to the mental workload for the arithmetic task, but not for the tracking task. On the contrary, the blink interval and heart variability revealed sensitivity to the workload for the tracking task, but not for the arithmetic task. Although no single measures revealed sufficient sensitivity, significant variation of mental workload was successfully detected for both tasks when all these measures were combined altogether.

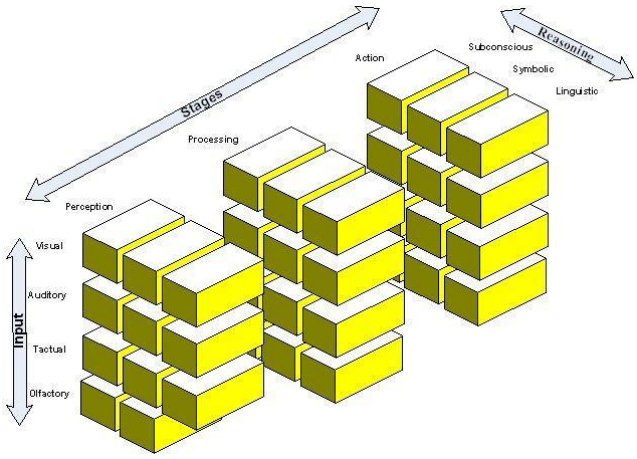


Figure 4. Multiple resource model (Wickens, 1984).

Klingner (2010) combined pupillometry and eye tracking in measuring cognitive load during visual tasks. In this study, fixation-aligned pupillary response averaging

method was developed. This method combines synchronized measurements of gaze direction and pupil size in order to assess short-term changes in cognitive load during unstructured visual tasks. Pupil measurements made during many instances of each task component can be aligned in time with respect to fixations and averaged, revealing any consistent pupillary response to that task component.

Chen et al. (2011) employed eight eye activity based features (e.g. features of eye blink, pupillary response and eye movement information) for real-time cognitive load measurement. An experiment using a computer-based training task showed that the three classes of eye features are capable of discriminating different cognitive load levels. Correlation analysis between various pairs of features suggests that significant improvements in discriminating different effort levels can be made by combining multiple features. Combined features of eye activity provide rich information on mental effort.

Consistent with the multiple resource theory (see Figure 4), previous studies indicated that task demands for different mental resource could be reflected by different physiological measures. The combination of multiple physiological measures has attracted increasing interests in cognitive workload studies, so that the explanatory power of multimodal information could be maximized.

13.2 Multimodal data fusion

The integration of multimodal information from multiple physiological measures is a non-trivial problem. Sometimes multiple measures could provide convergent results under single task condition, but inconsistent results under dual task condition. The way of data fusion is a key issue to efficiently and effectively integrate multimodal physiological features. For example, in a dual task experiment three workload measures based on brain activity, cardiac signal, and eye blink were combined into one composite measure using different weight coefficients (Ryu and Myung, 2005). It was shown that the composite measure significantly improved the sensitivity of workload assessment in the dual task.

Van Orden et al. (2001) employed artificial neural network to combine various eye activity based physiological features including blink frequency and duration, fixation frequency and time, saccadic extent, and pupil diameter for mental workload assessment. For each participant, a neural network model was trained on two sessions and tested on another session. Experimental results exhibited multiple eye activity based measures could be combined to produce reliable physiological index of workload in visuospatial task. In another experiment inducing fatigue (Van Orden et al., 2000), eye activity based features were also input to a neural network to estimate the fatigue state during the visual task performance.

Guhe et al. (2005) presented a Bayesian network approach to measure cognitive workload in real-time using multiple video based measures. The auditory two-back task, in which each participant was required to determine whether the current letter was equal to or different from the letter presented two back, was performed in the

experiment. Video based features including blink frequency, eye closure, saccadic movement, eye gaze, pupil dilatation, head movement, and mouth openness were recorded for each participant in the experiment. To make the model adaptive to both individual users and the specific task, Bayesian network was employed to fuse multiple video based measures for mental workload evaluation.

Zhang et al. (2004) proposed a machine learning approach for driver workload estimation using multiple physiological features including eye gaze and pupil diameter. Instead of analysing the significance of individual measures, all the measures were considered simultaneously during the task. The estimation of cognitive workload was optimized automatically with the use of machine learning techniques such as decision tree and Bayesian learning.

The combination of eye activity based physiological measures and facial skin temperature has also been proposed to enhance the sensitivity of mental workload measurement. Wang et al. (2007) presented a composite workload index based on facial skin temperature, eye blinks and pupil dilation. To improve the overall sensitivity to cognitive workload, the way of integrating eye activity features and facial skin temperature would be constructed through factor analysis and regression analysis.

14 Future work

Besides its sensitivity to changes in mental workload and usability as an objective measure, video based physiological measure has an attractive advantage that the measurement data can be captured in a non-intrusive and non-obtrusive way. The imaging sensors, especially the remote ones, minimize user interference and enable continuous data acquisition. Therefore, it is expected that video based physiological measures will become more and more popular in research and application areas involving cognitive workload. Meanwhile, various technique issues could be further investigated to improve the overall accuracy and sensitivity for mental workload assessment.

Video based workload measures such as pupillary response and skin temperature may be influenced by noisy factors relating to sensor technology. For example, subtle changes in physiological measures could be ignored due to the insufficient accuracy or resolution of the sensor. For remote eye tracker, the pupil area observed in video frames is also affected by the pose of human face. The sensitivity of physiological measures could be further enhanced by correcting the noises and distortions introduced during the sensing process.

As cognitive workload is multidimensional, single dimension of workload may have different effects on individual physiological measures. Previous studies also showed that different physiological measures could provide both overlapping and non-overlapping information about cognitive workload. Hence it will be useful to study the correlation between various video based physiological measures, especially under multitasking conditions.

Multiple physiological features could provide more information and result in better evaluation of mental workload than single physiological input. However, simple combination methods such as voting or linear weighting might not improve the overall accuracy and sensitivity for cognitive workload assessment. With the development of machine learning and information fusion techniques in recent years, probabilistic models and tools such as dynamic Bayesian network and Markov decision process could be employed to improve the fusion of multiple physiological measures.

On the other hand, an operator's mental workload during a task is determined by both demands of the task and capacity of the subject. From previous work on mental workload measures, it has been observed that physiological data is sensitive to the levels of task difficulty. Besides, the physiological measures may exhibit the effects of cognitive capacity as well. The correlation between video based physiological measures and subject's capacity should be further investigated to improve the explanatory power of physiological data.

Furthermore, both cognitive workload and physiological measures are influenced by many factors. For example, cognitive workload is dependent on operator's level of training, expertise, experience, motivation, etc. On the other hand, physiological measures are affected by various factors such as fatigue, stress, engagement, and environment. Ignoring these aspects may lead to the failure of physiological measures for mental workload assessment. The efficiency and effectiveness of video based physiological measures could be significantly enhanced when more of these factors are considered in a comprehensive way.

15 References

[1] U. Ahlstrom, J. Friedman-Berg: Using eye movement activity as a correlate of cognitive workload. International Journal of Aviation Psychology, vol. 36, pp. 623–636, 2006.

[2] R. Backs, A. Ryan, G. Wilson: Psychophysiological measures during continuous and manual performance. Human Factors, vol. 36, pp. 514–531, 1994.

[3] R. Backs, L. Walrath: Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied Ergonomics, vol. 23, pp. 243–254, 1992.

[4] B. Bailey, S. Iqbal: Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management, ACM Transactions on Computer-Human Interaction, vol. 14, pp. 21-1–21-28, 2008.

[5] B. Bailey, C. Busbey, S. Iqbal: TAPRAV: An interactive analysis tool for exploring workload aligned to models of task execution, Interacting with Computers, vol. 19, pp. 314–329, 2007.

[6] J. Beatty: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, vol. 91, pp. 276–292, 1982.

[7] A. Boehm-Davis, D. Gray, J. Schoelles: The eye blink as a physiological indicator of cognitive workload, Human Factors and Ergonomics Society Annual Meeting, pp. 116–119, 2006.

[8] J. Brookings, G. Wilson, C. Swain: Psychophysiological responses to changes in workload during simulated air traffic control, Biological Psychology, vol. 42, pp. 361–377, 1996.

[9] A. Byrne, R. Parasuraman: Psychophysiology and adaptive automation. Biological Psychology, vol. 42, pp. 249–268, 1996.

[10] B. Cain, A review of the mental workload literature, Technical Report, Defence Research and Development Canada Toronto, 2007.

[11] J. Callicott et al.: Physiological characteristics of capacity constraints in working memory as revealed by functional MRI, Cerebral Cortex, vol. 9, pp. 20-26, 1999.

[12] F. Chen: Designing Human Interface in Speech Technology, pp. 53–94, Springer, 2006.

[13] J. Coyne, C. Baldwin, A. Cole, C. Sibley, D. Roberts: Applying real time physiological measures of cognitive load to improve training, International Conference on Human-Computer Interaction, pp. 469–478, 2009.

[14] L. Di Stasi, V. Álvarez-Valbuena, J. Cañas, A. Maldonado, A. Catena, A. Antolí, A. Candido: Risk behaviour and mental workload: Multimodal assessment techniques applied to motorbike riding simulation, Transportation Research Part F: Traffic Psychology and Behaviour, vol. 12, pp. 361–370, 2009.

[15] L. Di Stasi, M. Marchitto, A. Antolí, T. Baccino, J. Cañas, Approximation of on-line mental workload index in ATC simulated multitasks, Journal of Air Transport Management, in press, 2010.

[16] C. Fogarty, J. Stern: Eye movements and blinks: Their relationship to higher cognitive processes. International Journal of Psychophysiology, vol. 8, pp. 35–42, 1989.

[17] L. Fournier, G. Wilson, C. Swain: Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training, International Journal of Psychophysiology, vol. 31, pp. 129–145, 1999.

[18] H. Genno, K. Ishikawa, O. Kanbara, M. Kikumoto, Y. Fujiwara, R. Suzuki, M. Osumi: Using facial skin temperature to objectively evaluate sensations. International Journal of Industrial Ergonomics, vol. 19, pp. 161–171, 1997.

[19] R. Gingell, Review of workload measurement, analysis and interpretation methods, Technical Report, European Organisation for the Safety of Air Navigation, 2003.

[20] D. Gopher, E. Donchin: Workload-An examination of the concept. K. Boff, L. Kaufman, J. Thomas, (eds.) Handbook of Perception and Human Performance, Wiley, 1986.

[21] T. Greef, H. Lafeber, H. Oostendorp, J. Lindenberg: Eye movement as indicators of mental workload to trigger adaptive automation, International Conference on Human-Computer Interaction, pp. 219–228, 2009.

[22] M. Grootjen, M. Neerincx, J. Weert: Task-based interpretation of operator state information for adaptive support. D. Schmorrow, M. Stanney, M. Reeves, (eds.) Foundations of Augmented Cognition (2nd edn.), pp. 236–242, 2006.

[23] M. Grootjen, M. Neerincx, J. Weert, and K. Truong: Measuring cognitive task load on a naval ship: implications of a real world environment, International Conference on Human-Computer Interaction, pp. 147–156, 2007.

[24] M. Guhe, W. Liao, Z. Zhu, Q. Ji, D. Gray, J. Schoelles: Non-intrusive measurement of workload in real-time. Human Factors and Ergonomics Society Annual Meeting, pp. 1157–1161, 2005.

[25] T. Hankins, G. Wilson: A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviation, Space, and Environmental Medicine, vol. 69, pp. 360–367, 1998.

[26] S. Hart, L. Staveland: Development of the NASA task load index (TLX): Results of experimental and theoretical research. P. Hancock, N. Meshkati (eds.): Human Workload, pp. 138–183, North-Holland, 1988.

[27] P. He, B. Yang, S. Hubbard, J. Estepp, G. Wilson: A sensor positioning system for functional near-infrared neuroimaging. International Conference on Human-Computer Interaction, pp. 30–37, 2007.

[28] R. Heishman, Z. Duric: Using eye blinks as a tool for augmented cognition. International Conference on Human-Computer Interaction, pp. 84–93, 2007.

[29] S. Iqbal, X. Zheng, B. Bailey: Task-evoked pupillary response to mental workload in human-computer interaction. ACM Conference on Human Factors in Computing Systems (CHI), pp. 1477–1480, 2004.

[30] M. Izzetoglu, K. Izzetoglu, S. Bunce, H. Ayaz, A. Devaraj, B. Onaral, K. Pourrezaei: Functional near-infrared neuroimaging. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 13, pp. 153–159, 2005.

[31] M. Just, P. Carpenter, A. Miyake: Neuroindices of cognitive workload neuroimaging, pupillometric and event-related potential studies of brain work, Theoretical Issues in Ergonomics Science, vol. 4, pp. 56–88, 2003.

[32] J. Klingner, R. Kumar, P. Hanrahan: Measuring the task-evoked pupillary response with a remote eye tracker. Eye Tracking Research and Applications Symposium, pp. 69–72, 2008.

[33] A. Kramer: Physiological metrics of mental workload: A review of recent progress, D. Damos (ed.), Multiple-Task Performance, pp. 279–328, Taylor and Francis, 1991.

[34] T. Lin, A. Imamiya: Evaluating usability based on multimodal information: An empirical study. International Conference on Multimodal Interfaces, pp. 364–371, 2006.

[35] P. Marshall: The index of cognitive activity: Measuring cognitive workload. IEEE Human Factors Meeting, pp. 7-5–7-9, 2002.

[36] J. May, R. Kennedy, M. Williams, W. Dunlap, J. Brannan: Eye movement indices of mental workload. Acta Psychologica, vol. 75, pp. 75–89, 1990.

[37] A. Murata, H. Iwase: Evaluation of mental workload by fluctuation analysis of pupil area. Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 3094–3097, 1998.

[38] M. Nakayama, Y. Shimizu: Frequency analysis of task evoked pupillary response and eye-movement, Eye Tracking Research and Applications Symposium, pp. 71–76, 2004.

[39] D. Neumann: Effect of varying levels of mental workload on startle eyeblink modulation. Ergonomics, vol. 45, pp. 583–602, 2002.

[40] R. O'Donnell, F. Eggemeier: Workload assessment methodology. K. Boff, L. Kaufman, J. Thomas (eds.): Handbook of Perception and Human Performance. Cognitive Processes and Performance, vol. 2, pp. 42-1–42-49, Wiley, 1986.

[41] K. Or, G. Duffy: Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. Occupational Ergonomics, vol. 7, pp. 83–94, 2007.

[42] F. Paas, E. Juhani, H. Tabbers, P. Van Gerven: Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist, vol. 38, pp. 63–71, 2003.

[43] O. Palinko, A. Kun, A. Shyrokov, P. Heeman: Estimating cognitive load using remote eye tracking in a driving simulator, Eye Tracking Research and Applications Symposium, pp. 141–144, 2010.

[44] I. Pavlidis, J. Levine, P. Baukol: Thermal imaging for anxiety detection. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, pp. 104–109, 2000.

[45] M. Pomplun, S. Sunkara: Pupil dilation as an indicator of cognitive workload in human-computer interaction. International Conference on Human-Computer Interaction, pp. 542–546, 2003.

[46] L. Prinzel, F. Freeman, M. Scerbo, P. Mikulka, A. Pope: A closed-loop system for examining psychophysiological measures for adaptive task allocation. International Journal of Aviation Psychology, pp. 393–410, 1999.

[47] C. Puri, L. Olson, I. Pavlidis, J. Levine, J. Starren: StressCam: Non-contact measurement of users' emotional states through thermal imaging. ACM conference on Human factors in computing systems (CHI), pp. 1725–1728, 2005.

[48] M. Recarte, E. Perez, A. Conchillo, L. Nunes: Mental workload and visual impairment: differences between pupil, blink, and subjective rating. Spanish Journal of Psychology, vol. 11, pp. 374–385, 2008.

[49] K. Ryu, R. Myung: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. International Journal of Industrial Ergonomics, vol. 35, pp. 991–1009, 2005.

[50] M. Scerbo, F. Freeman, P. Mikula, R. Parasuraman, F. Di Nocero, L. Prinzel: The efficacy of psychophysiological measures for implementing adaptive technology. Technical Report, NASA Langley Research Center, 2001.

[51] M. Schwalm, A. Keinath, H. Zimmer: Pupillometry as a method for measuring mental workload with a simulated driving task, D. Waard, F. Flemisch, B. Lorenz, H. Oberheid, K. Brookhuis (eds.), Human Factors for Assistance and Automation, pp. 75–87, 2008.

[52] L. Sciarini, D. Nicholson: Assessing cognitive state with multiple physiological measures: A modular approach, International Conference on Human-Computer Interaction, pp. 533–542, 2009.

[53] J. Sirevaag, A. Stern: Ocular measures of fatigue and cognitive factors. W. Backs, W. Boucsein (eds.): Engineering Psychophysiology - Issues and applications, pp. 269–287, 2000.

[54] J. Sweller: Cognitive load during problem solving: Effects on learning. Cognitive Science: A Multidisciplinary Journal, vol. 12, pp. 257–285, 1988.

[55] N. Thomas, Y. Du, T. Artavatkun, J. She: Non-intrusive personalized mental workload evaluation for exercise intensity measure, International Conference on Human-Computer Interaction, pp. 315–322, 2009.

[56] Y. Tsai, E. Viirre, C. Strychacz, B. Chase, T. Jung: Task performance and eye activity - predicting behavior relating to cognitive workload. Aviation, Space, and Environmental Medicine, vol. 78, pp. B176–B185, 2007.

[57] M. Tungare, M. Perez-Quinones: Mental workload in multi-device personal information management, ACM Conference on Human Factors in Computing Systems (CHI), pp. 3431–3436, 2009.

[58] P. Van Gerven, F. Paas, J. Van Merriënboer, H. Schmidt: Memory load and the cognitive pupillary response in aging, Psychophysiology, vol. 41, pp. 167–174, 2003.

[59] K. Van Orden, T. Jung, S. Makeig: Combined eye activity measures accurately estimate changes in sustained visual task performance. Biological Psychology, vol. 52, pp. 221–240, 2000.

[60] K. Van Orden, W. Limbert, S. Makeig, T. Jung.: Eye activity correlates of workload during visuospatial memory task. Human factors, vol. 43, pp. 111–121, 2001.

[61] H. Veltman, W. Vos: Facial temperature as a measure of operator state. International Conference on Augmented Cognition, pp. 22–27, 2005.

[62] J. Veltman, A. Gaillard: Physiological workload reactions to increasing levels of task difficulty. Ergonomics, vol. 41, pp. 656–669, 1998.

[63] J. Veltman, C. Jansen: The role of operator state assessment in adaptive automation, Technical Report, TNO Human Factors Research Institute, 2006.

[64] S. Verney, E. Granholm, D. Dionisio: Pupillary responses and processing resources on the visual backward masking task. Psychophysiology, vol. 38, pp. 76–83, 2001.

[65] J. Voskamp, B. Urban: Measuring cognitive workload in non-military scenarios criteria for sensor technologies, International Conference on Human-Computer Interaction, pp. 304–310, 2009.

[66] L.-M. Wang, V. Duffy, Y. Du: A composite measure for the evaluation of mental workload, International Conference on Human-Computer Interaction, pp. 460–466, 2007.

[67] C. Wickens: Processing resources in attention. R. Parasuraman, D. Davies, (eds.) Varieties of Attention, pp. 63–102. Academic Press, 1984.

[68] C. Wickens: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, vol. 3, pp. 150–177, 2002.

[69] W. Wierwille, F. Eggemeier: Recommendations for mental workload measurements in a test and evaluation environment. Human Factors, vol. 35, pp. 263–281, 1993.

[70] G. Wilson: An analysis of mental workload in pilots during flight using multiple psychophysiological measures. International Journal of Aviation Psychology, vol. 12, pp. 3–18, 2002.

[71] G. Wilson et al.: Operator functional state assessment. Technical Report, Research and Technology Organisation, North Atlantic Treaty Organisation (NATO), 2004.

[72] G. Wilson, C. Russel: Psychophysiologically versus task determined adaptive aiding accomplishment. D. Schmorrow, K. Stanney, L. Reeves (eds.): Foundations of Augmented Cognition (2nd edn.), pp. 201–207, 2006.

[73] B. Yin, F. Chen, N. Ruiz, E. Ambikairajah: Speech-based Cognitive Load Monitoring System, IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 2041–2044, 2008.

[74] C. Zhang, C. Zheng, X. Yu: Evaluation of mental fatigue based on multipsychophysiological parameters and kernel learning algorithms, Chinese Science Bulletin, vol. 53, pp. 1835–1847, 2008.

[75] Y. Zhang, Y. Owechko, J. Zhang: Driver cognitive workload estimation - a data driven perspective. International IEEE Conference on Intelligent Transportation Systems, pp. 642–647, 2004.

[76] M. Bartels, S. Marshall: Measuring Cognitive Workload Across Different Eye Tracking Hardware Platforms. Proceedings of the Symposium on Eye Tracking Research and Applications Symposium, pp. 161-164, 2012.

[77] L. Richstone, M. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, L. Kavoussi: Eye metrics as an objective assessment of surgical skill. Annals of Surgery. Jul, vol. 252, no. 1, pp. 177-82, 2010.

[78] B. Rebsamen, T. B. Penney, K. Kwok: EEG-based measure of cognitive workload during a mental arithmetic task. Communications in Computer and Information Science, 174 CCIS (PART 2), pp. 304-307, 2011.

[79] B. Rebsamen, K. Kwok, T. B. Penney: Evaluation Of Cognitive Workload From EEG During A Mental Arithmetic Task. In Proceedings of the 55th Annual Meeting of the Human Factors and Ergonomics Society, vol. 52, no. 1, pp.1342-1345, 2011.

[80] G. Sammer: Functional Magnetic Resonance Imaging (fMRI) and Workload Assessment. In International Encyclopedia of Ergonomics and Human Factors, Second Edition, Edited by W. Karwowski, Chapter 610, CRC Press, 2006.

[81] B.B. Zheng, X. X. Jiang, G. G. Tien, A. A. Meneghetti, O. N. M. ON Panton, M. S. MS Atkins: Workload assessment of surgeons: correlation between NASA TLX and blinks. Surgical Endoscopy, vol. 26, no. 10, pp. 2746-2750, 2012.

[82] S. E. Kerick and L.E. Allender: Effects of cognitive workload on decision accuracy, shooting performance, and cortical activity of soldiers. In Proceedings of 5th International Conference on Psychophysiology In Ergonomics, 2004.

[83] A. Berthold and A. Jameson: Interpreting Symptoms of Cognitive Load in Speech Input. UM99, 1999.

[84] M.A. Khawaja, N. Ruiz, and F. Chen: Think before you talk: An empirical study of Relationship between speech pauses and cognitive load. Paper presented at the

Australasian Computer-Human Interaction Conference (OzCHI'08), Cairns, Australia, 2008.

[85] M.A. Khawaja, F. Chen, C. Owen, and G. Hickey: Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design. In Proceedings of the 12th IFIP International Conference on Human-Computer Interaction (INTERACT'09), Uppsala, Sweden, pp. 485-489, August 2009.

[86] V. Demberg and A. Sayeed: Linguistic cognitive load: implications for automotive UIs. In Proceedings of AutomotiveUI'11, Salzburg, Austria, November 29-December 2, 2011.

[87] M.A. Khawaja, F. Chen, and N. Marcus: Analysis of Collaborative Communication for Linguistic Cues of Cognitive Load. International Journal of Human Factors and Ergonomic Society, vol. 54, no 4. pp 518-529, August 2012.

[88] C. Mueller, B. Großmann-Hutter, A. Jameson, R. Rummer, and F. Wittig: Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. UM2001, 2001.

[89] A. Jameson, J. Kiefer, C. Mueller, B. Großmann-Hutter, F. Wittig, and R. Rummer: Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech. Journal of Computer Science and Technology, 2006.

[90] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah: Speech-based cognitive load monitoring system. Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Las Vegas, 2041-2044, March/April 2008.

[91] P. Le, E. Ambikairajah, J. Epps, S. Vidhyasaharan, and E. Choi: Investigation of spectral centroid features for cognitive load classification. Speech Communication, vol. 53, no. 4, pp. 540-551, April 2011.

[92] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah: Glottal features for speech-based cognitive load classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10), pp. 5234–5237, 2010.

[93] T.F. Yap, J. Epps, E. Ambikairajah, and E. Choi: Voice Source Features for Cognitive Load Classification. In Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'11), Prague, Czech Republic, pp. 5700-5703, May 2011a.

[94] T. F. Yap, J. Epps, E. Ambikairajah, and E. Choi: Formant frequencies under cognitive load: Effects and classification. EURASIP Journal on Advances in Signal Processing, vol. 2011, Article ID 219253, 11 pages, January 2011b.

[95] P. Le, J. Epps, E. Choi, and E. Ambikairajah: A study of voice source and vocal tract filter based features in cognitive load classification. In Proceedings of International Conference on Pattern Recognition, pp. 4516–4519, 2010.

[96] C. M. Schulz, E. Schneider, L. Fritz, J. Vockeroth, A. Hapfelmeier, M. Wasmaier, E. F. Kochs, and G. Schneider: Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. British Journal of Anaesthesia, vol 106, no 1, pp. 44-50, 2011.

[97] J. Xu, Y. Wang, F. Chen, E. Choi, G. Li, S. Chen, and S. Hussain: Pupillary Response Based Cognitive Workload Index under Luminance and Emotional Changes. In Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI'11), Vancouver, Canada, pp. 1627-1632, 2011a.

[98] J. Xu, Y. Wang, F. Chen, and E. Choi: Pupillary Response Based Cognitive Workload Measurement under Luminance Changes. In Proceedings of IFIP International Conference on Human-Computer Interaction (INTERACT'11), Lisbon, Portugal, pp. 178-185, 2011b.

[99] A. Ganguly: Surprise and Cognitive Workload Effect on Pupil Size. http://avik-ganguly.blogspot.com.au/2012/04/surprise-and-cognitive-workload-effect.html, April 2012.

[100] O. Hasan: Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. Expert Systems with Applications, vol. 36, pp. 2027-2036, 2009.

[101] P.F. Diez, E. Laciaret et al.: A comparative study of the performance of different spectral estimation methods for classification of mental tasks. In Proceedings of the 30th Int' IEEE EMBS Conference, pp.1155-1158, 2008.

[102] P. Zarjam, J. Epps, and F. Chen: Characterizing Working Memory Load Using EEG Delta Activity. In Proceedings of the 19th European Signal Processing Conference (EUSIPCO'11), Barcelona, Spain, pp. 1554-1558, August 2011.

[103] P. Zarjam, J. Epps, and F. Chen: Spectral EEG Features for Evaluating Cognitive Load. In Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'11), Boston, USA, pp. 3841-3844, August 2011.

[104] E.W. Anderson, K.C. Potter, L.E. Matzen, et al.: A User Study of Visualization Effectiveness Using EEG and Cognitive Load. Computer Graphics Forum, vol 30, no 3, pp. 791-800, 2011.

[105] P. Zarjam, J. Epps, and N. H. Lovell: Characterizing mental load in an arithmetic task using entropy-based features. In Proceedings of the 11th International Conference on Information Science, Signal Processing and their applications (ISSPA'2012), pp. 199 – 204, 2012a.

[106] P. Zarjam, J. Epps, N. H. Lovell, and F. Chen: Characterization of Memory Load in an Arithmetic Task using Non-Linear Analysis of EEG Signals. In Proceedings of the 34th IEEE Engineering in Medicine and Biology Conference (EMBC'2012), pp. 3519-3522, California, USA, 2012b.

[107] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen: Galvanic Skin Response (GSR) as an Index of Cognitive Load. In Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI'07), San Jose, pp. 2651-2656, April/May 2007.

[108] E. Haapalainen, S.J. Kim, J.F. Forlizzi, and A.K. Dey: Psycho-Physiological Measures for Assessing Cognitive Load. In Proceedings of Ubicomp 2010, pp.301-310, 2010.

[109] J. Son and M. Park: Estimating Cognitive Load Complexity Using Performanceand Physiological Data in a Driving Simulator. In Proceedings of AutomotiveUI'11, Salzburg, Austria, November 29-December 2, 2011.

[110] N. Nourbakhsh, Y. Wang, F. Chen and R. Calvo: Using Galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In Proceedings of Australian Computer-Human Interaction Conference (OzCHI), 2012.

[111] N. Ruiz, R. Taib, Y. Shi, E. Choi, and F. Chen: Using Pen Input Features as Indices of Cognitive Load. In Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI'07), Nagoya, Japan, pp. 315-318, Nov. 2007.

[112] K. Yu, J. Epps, and F. Chen: Cognitive Load Evaluation with Pen Orientation and Pressure. In Proceedings of ICMI Workshop on Inferring Cognitive and Emotional States from Multimodal Measures (MMCogEmS'11), Alicante, Spain, November 2011.

[113] A.A, Zekveld, S.E. Kramer, and J.M. Festen: Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. Ear Hear, vol. 32, no 4, pp. 498-510, 2011.

[114] S. Chen, J. Epps, N. Ruiz, and F. Chen: Eye Activity as a Measure of Human Mental Effort in HCI. In Proceedings of International Conference on Intelligent User Interfaces (IUI'11), Palo, Alto, U.S.A., pp. 315-318, 2011.

[115] J. Klingner, B. Tversky, and P. Hanrahan: Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. Psychophysiology, vol. 48, no 3, pp. 323-32, 2011.

[116] J. Klinger: Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking. PhD Thesis, Department of Computer Science, Stanford University, USA, 2010.

Attachment B

Classification of Working Memory Load Using Wavelet Complexity Features of EEG Signals

Pega Zarjam[1,2], Julien Epps[1,2], Fang Chen[2], and Nigel H. Lovell[1,3]

[1] School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
p.zarjam@student.unsw.edu.au,
{j.epps,n.lovell}@unsw.edu.au
[2] ATP Research Laboratory, National ICT Australia, Eveleigh, NSW 2015, Australia
fang.chen@nicta.com
[3] Graduate School of Biomedical Eng. The University of NSW, Sydney, NSW 2052, Australia

Abstract. We investigate the use of wavelet-based complexity measures of electroencephalogram (EEG) signals to evaluate changes in working memory load during the performance of a cognitive task with varying difficulty/load levels. Extracted wavelet-complexity measures associated with four entropic measures; that is Shannon, Tsallis, Escort-Tsallis and Renyi entropies demonstrate good discrimination among seven load levels imposed on the working memory with a classification rate of up to 96% using signals recorded from the frontal lobe of the brain. The extracted measures' values show a consistent decrease in the selected channels in two frontal and occipital lobes, as the memory load increases, indicating the EEGs disorder declines while the complexity grows. This illustrates that the brain behaves in a more organized manner characterized by more order and maximal complexity when dealing with higher load levels. The growing complexity can also reflect the higher activation of neural networks involved, as the task load increases.

1 Introduction

The electroencephalogram (EEG) is a non-invasive neuroimaging technique widely used for the diagnosis of neurological dysfunctions and the understanding of cognitive processes. Practically, it can be a very effective apparatus for the understanding of the complex behavior of the brain in different cognitive states due to its high temporal resolution, relative ease of use, and a comparably low cost [1]. Each cognitive process activates local and spatial cortical networks to an extent depending on task specificity and complexity [2].

Measuring the amount of cognitive/working memory load when performing a cognitive process is of high importance for the prevention of decision-making errors, and the development of adaptive user interfaces [3]. This is necessary to avoid memory overload and maintain efficiency and productivity during tasks, especially in critical/high mental load workplaces such as persons working in the areas of air traffic control, military operations and emergency/interventional medicine.

T. Huang et al. (Eds.): ICONIP 2012, Part II, LNCS 7664, pp. 692–699, 2012.

Currently, different methods are available to measure working memory load, such as; behavioral/physiological techniques or performance-based/subjective ratings methods. Among them, EEG has been rated as the best physiological method, offering more reliability and sensitivity, when measuring memory load [4].

A range of features; mainly power spectral-based, have been applied for measuring the working memory load using EEG signals, previously [5-7]. The application of non-linear/dynamical measures in classifying different mental tasks or the comparison with the rest condition is more recent, and measures like correlation dimension (CD) [8, 9], Hurst exponent (HE), approximate entropy (ApEn) and largest Lyapunov exponent (LLE) [10, 11] have been used to measure the complexity or irregularity of the underlying brain dynamics. In [10], it is concluded that the brain reflects a lesser degree of cognitive activity (shown by less correlation dimension/complexity) when the participants are subject to sound or reflexologic stimulation compared with the normal state.

Since dynamical features had not been used in the study of measuring memory load previously and also the question of whether the complexity or order/regularity of the EEG signals change when the imposed load varies, the authors aimed at addressing these questions in [12, 13]. In these studies, features such as: spectral entropy, CD, HE, and ApEn proved to be a good discriminator of imposed memory load and indicator of higher predictability and less irregularity/more order in the brain activity when dealing with higher memory load. CD feature also showed that the brain activity dimension/complexity increases with the increase of memory load. However, in our previous studies, the relationship between the signals' order/regularity and its complexity was not explicitly investigated. In this study, we investigate not only a recently proposed feature set; based on wavelet-complexity measures [14-16], for discriminating the memory load, but also the signals' changing complexity and order relationship with varying memory load imposed, and their implication on the neural activations towards a better understanding of the brain dynamics when dealing with higher loads.

2 Materials and Methods

2.1 Experiment and Dataset

EEG signals were acquired from twelve healthy male subjects; postgraduate students aged between 24-30 years. In the experiment, the participants were asked to do an arithmetic task (an addition task with varying difficulty level).

Each time, the numbers to be added were displayed sequentially and in Arabic notation, on a laptop PC with a viewing distance of 70 cm to the subject. The difficulty level was manipulated by varying the n-digit numbers used and carries required to calculate the addition, as follows: in very low level (L1); 1&2 digit numbers with no carry, in low level (L2); carry is introduced to L1, in medium level (L3); 2 digit numbers with one carry, in medium-high level (L4); 2 digit numbers with two carries, in high level (L5); 2&3 digit numbers with one carry, in very high level (L6); 2&3 digit numbers with two carries, in extremely high level (L7); 3 digit numbers with three carries. The subjects were required to click on the correct answer using the mouse left button, using the minimum possible finger movement. In the baseline/rest condition,

conducted after the experiment, the participants were asked to sit relaxed and keep their eyes closed. To minimize any muscle movement artifact (EMG) during the recording, the subjects were asked to avoid any unnecessary physical movements and their hand was placed in a fixed position.

The subjects' EEG signals were recorded using an Active Two system. Each recording contained 32 EEG channels mounted in an elastic cap, according to the extended international 10 - 20 system. A linked earlobe reference was used and impedance was kept under 5 kÙ. The EEG signals were passed through a band-pass filter with cut-off frequencies of $0.1 - 100$ Hz and were recorded at a $f_s = 256$ Hz sampling rate. To select the epochs which contained minimal EMG artifact, each recording was judged by visual inspection. As a result, 70 seconds (out of 90 seconds of each task level recording) for each subject was considered. This portion of the recordings included EOG and ECG artifacts, which were not removed.

2.2 EEG Source Localization

Source localization can be used to estimate the localization and distribution of electrical events in brain disorders [17]. We used this technique to narrow down the number of channels under study and select discriminatory channels, as described in our previous work [12].

2.3 Wavelet-Based Complexity Measures

In studying EEG signals, entropy is a measure of order and more specifically, a degree of synchrony of the cell groups contributed in different neural responses [18]. If this entropy is considered with the system's likely state/architecture, one can define system complexity as a form of statistical complexity measure [16].

General form of wavelet statistical complexity measures can be found in [16], which uses different entropy types and distance measures. In this study, we use the complexity measure of $C_q^{(k)}[P]$ given in (1), which is based on the Kullback/q-Kullback distance measure [16], as below:

$$\mathrm{C_q^{(k)}[P] = (1 - H_q^{(k)}[P]).\, H_q^{(k)}[P]; k = 1,2,3,4} \tag{1}$$

In (1), P is the probability distribution of the Discrete Wavelet Transform (DWT) of parameter under study, q is the entropic index $(0 \le q \le 1)$ and k refers to the entropy types used as follows [18]:

$$\text{Shannon: } \mathrm{H_1^{(1)}[P] = H_{SH} = -\sum_{i=1}^{N} p_i \ln (p_i)} \tag{2}$$

$$\text{Tsallis: } \mathrm{H_q^{(2)}[P] = H_{TS} = \frac{1}{q-1}\sum_{i=1}^{N}[(p_i - (p_i)^q]} \tag{3}$$

$$\text{Escort-Tsallis: } \mathrm{H_q^{(3)}[P] = H_{ETS} = \frac{1}{q-1}\left(1 - \left[\sum_{i=1}^{N}(p_i)^{1/q}\right]^{-q}\right)} \tag{4}$$

$$\text{Renyi: } \mathrm{H_q^{(4)}[P] = H_{RE} = \frac{1}{1-q}\ln\left[\sum_{i=1}^{N}(p_i)^q\right]} \tag{5}$$

where p_i is the distribution of the DWT parameter of the under study EEG segment (i^{th}) and $q = 1$ for Shannon entropy and $0 \leq q < 1$ for other entropies.

3 Experimental Results

Our earlier source localization results demonstrated that mainly the frontal and occipital regions of the brain were the most influenced regions, in all the task load levels across all twelve subjects ([12, 13]). Therefore, only EEG channels located in these two regions (i.e. the frontal channels Fp1, AF3, F7, F3, FC1, FC5 FC6, FC2, F4, F8, AF4, Fp2 and the occipital channels PO3, O1, Oz, O2, PO4) were considered for further analysis.

We decomposed the EEG signals of length $T = 5$ seconds (non-overlapping), into five levels (scales) using Daubechies-4 mother wavelet. We denote the under study wavelet parameter here are wavelet coefficients. For instance, in case of approximate coefficients at the 5th level (which corresponds to the delta frequency band) we have:

$$a_5 = [a_{51}\ a_{52} \ldots\ a_{5N}] \tag{6}$$

where $N = 40$ is the number of approximate coefficients at the 5th level; ($N = \frac{Tfs}{2^5} = 40$). P in equations (2)-(5) is therefore defined as:

$$P = \frac{a_5}{\sum_{i=1}^{N} a_{5i}} \tag{7}$$

Then, we calculated four entropic features; H_{SH}, H_{TS}, H_{ETS} and H_{RE} using equations (2)-(5) for each EEG segment. The index q in H_{TS}, H_{ETS} and $H_{RE}$was varied to find its optimal value for the purpose of the load discrimination. The feature values showed a decreasing trend as the task load increased in many channels of interest. For instance, the extracted H_{RE} values for channel Fp1 of subject 1 for three load levels are L1=871.77, L4= 865.61, and L7= 859.68, while for the rest condition=877.70.

For illustration purposes, Fig. 1 shows the median of the extracted H_{RE} from the frontal channels in scale 5, for channel F7 of subject 1, for two extreme values of q; (a) $q = 0.9$, (b) $q = 0.1$, in the delta frequency band. As shown, the median of the extracted H_{RE} are able to distinguish the seven task loads better with q closer to 1, as it consistently reveals a decreasing median with increasing task load.

Following preliminary analysis, those features and frequency bands which show a consistent decreasing trend with increasing load across all twelve subjects, are summarized as follows: for the frontal lobe; channels Fp1, F7, F3, FC5, FC6, FC2, and AF4 in the delta band, channels FC5, AF4 in the alpha band; for the occipital lobe; channels PO3, O1, and O2, in the delta band. For illustration purposes, Fig. 2(a) shows the median of the extracted H_{RE} from the frontal channels in scale 5, across all subjects. We then calculated the complexity values for each entropic feature, using (1). The results showed that the complexity values increases as the task load increases, in the above selected channels. For illustration purposes, the complexity values corresponding to Fig. 1(a) for channel F7 of subject 1, using H_{RE} entropy is shown in Fig. 2(b). This demonstrates that the signal complexity increases with increasing task load, while the corresponding signal entropy/disorder decreases in Fig. 1(a).

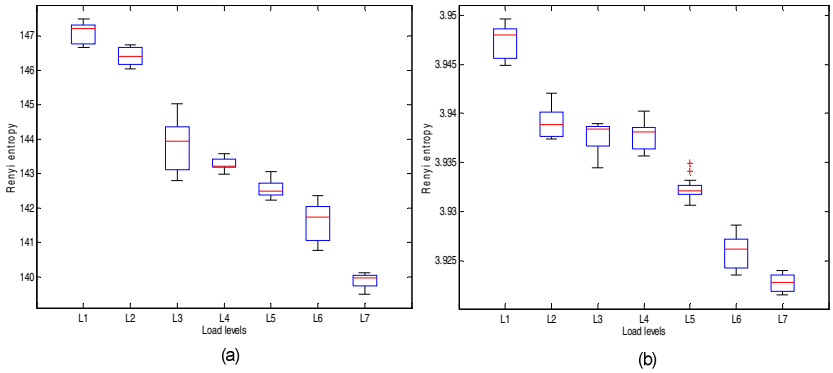


Fig. 1. The Renyi entropy variations for (a) $q = 0.9$, (b) $q = 0.1$ with the load levels, for channel F7 of subject 1. On each box, the red mark is the median; the edges of the box are the 25th and the 75th percentiles.

In order to study the performance of the entropic features in classifying different load levels, we applied the four extracted features from the EEG segments acquired from the selected channels into an Artificial Neural Network (ANN) classifier. Based on experimental results, we chose a multi-layer perceptron ANN, with a first hidden layer of 20 neurons, a second hidden layer of 14 neurons and an output layer of 7 neurons corresponding to 7 load levels. 75% of the data (for each task level for twelve subjects) were used for training and the remainder for testing, in a subject-dependent arrangement. Since the delta band contained more selected channels for all the extracted features across all the subjects, we considered the classification accuracy of the features only in this frequency band. The classification results are summarized in Table 1.

Table 1. Classification accuracy of the four entropic measures ($q = 1$ for Shannon and $q = 0.9$, for the remaining entropies) extracted from the delta band from channels in the two identified regions of interest

Channels	Feature	Accuracy %
Frontal: Fp1, F7, F3, FC5, FC6, FC2, and AF4	H_{SH}	96.83
	H_{TS}	94.18
	H_{ETS}	82.10
	H_{RE}	89.42
Occipital: PO3, O1, and O2	H_{SH}	85.71
	H_{TS}	88.36
	H_{ETS}	51.32
	H_{RE}	83.60

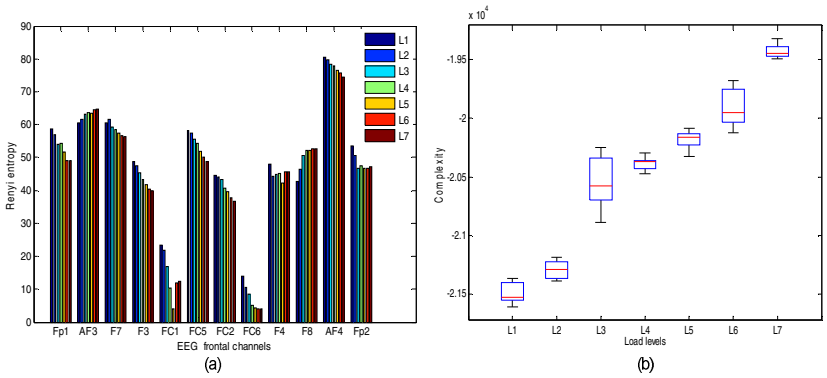


Fig. 2. (a) Medians of the Renyi entropy extracted from segmented EEG data in the delta band, from the frontal lobe, across twelve subjects. (b) The complexity variations with the load level increase from L1 to L7 for channel F7, using extracted H_{RE}, for subject 1.

4 Discussion

In this study, we investigated the use of four entropic measures in different wavelet levels (wavelet-complexity features) for discriminating working memory load in a cognitive task with seven load levels. The extracted measures from the selected channels; picked up by source localization from the frontal and occipital lobes of the brain, were found to be successful in memory load discrimination. The decline in the median values of the entropy features as the task load increased demonstrates that the degree of the disorder decreases as the task load/working memory load imposed increases.

The complexity values measured by each entropic measure showed an increasing trend as the task load increased. This indicates that with increasing memory load, not only the disorder of the signals declines but also the complexity grows. This can demonstrate a more organized manner of the brain characterized by more order and maximal complexity at the same time, when dealing with higher load levels. Practically, more order implicates higher degree of synchrony of the cell groups contributed in neural responses [18] and more complexity indicates higher activation of the neurons. This can confirm the changing dynamics of the brain when performing a task with different load/difficulty levels. This is supported by [8], in which the complexity of EEG signals (shown by correlation dimension) increases as more difficult cognitive tasks are performed and it indicates the level of vigilance and mental activity. This is also confirmed by previous studies that the increasing workload is reflected by more activity and mostly in the frontal lobe of the brain [19, 20]. On the other hand, our classification results revealed that the extracted features show a significantly higher accuracy for the selected frontal channels compared with the selected occipital channels.

We also examined different values of entropic index of q to find the optimal value for the purpose of task load discrimination in this study. The results showed the larger the value of q (closer to 1) the better the different load levels were distinguished, for

the three measures of Tsallis, Escort-Tsallis and Renyi. This reason could be that as q increases the three entropic measures become closer to Shannon entropy, for which the classification rate outperformed the rest of the features in the frontal channels. Its classification accuracy is closely followed by Tsallis entropy which is a generalisation of Shannon entropy.

Since the used complexity formula is based on entropy, one may criticise that it could carry the same information as entropy. But in [21], it is demonstrated that this simple entropy-based measure is really an indicator of complexity in many systems.

The frequency band analysis showed that the delta is the most promising band for task load discrimination, including more selected channels for the four measures in our study. This is while, only two channels in the alpha band and no channel in the theta band, showed significant discrimination among all seven load levels. This was confirmed by classification results, as well. This is in line with previous studies showing that the delta activity could be an indicator of attention during some mental tasks, so that by increasing task demand, participant's attention to the task and also the delta band activity increases [22].

Comparison of the rest condition signals, recorded after task accomplishment, with the task condition signals showed that the entropy value of the highest load level is lower than the rest condition in all the subjects. This can indicate that the brain is in a less disordered (more ordered/focused) state when conducting a cognitive task.

The entropic features not only add to the collection of suitable feature sets for characterizing working memory load previously applied by the authors, but also proved to be computationally more efficient than using non-linear dynamical features such as correlation dimension, approximate entropy and Hurst exponent. Furthermore, the entropic features are relatively free of parameter tuning which is critical and highly application-dependent for non-linear dynamical features.

For future work, this method could be validated on a larger database and in more realistic environments and conducting other cognitive tasks with a focus on cognitive overload.

Acknowledgements. This research was supported by the Asian Office of Aerospace Research & Development, Grant No. FA2386-12-1-4049.

References

1. Zander, O.T., Kothe, C.: Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general. Journal of Neural Eng. 8, 1–5 (2011)
2. Dimitriadis, S.I., Laskaris, N.A., Tsirka, V., Vourkas, M., Micheloyannis, S.: What does delta band tell us about cognitive processes: A mental calculation study. Neuroscience Letters 483, 11–15 (2010)
3. Paas, F., Tuovinen, J.E., Tabbers, H., Gerven, P.W.M.V.: Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist 38, 63–71 (2003)
4. Antonenko, P., Paas, F., Grabner, R., Gog, T.V.: Using Electroencephalography to measure cognitive load. Educational Psychology Review 22, 425–438 (2010)
5. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. Brain Research Reviews 29, 169–195 (1999)

6. Spencer, K.M., Polich, J.: Poststimulus EEG spectral analysis and P300: Attention, task, and probability. Psychophysiology 36, 220–232 (1999)
7. Diez, P.F., Laciar, E., Mut, V., Avila, E., Torres, A.: A comparative study of the performance of different spectral estimation methods for classification of mental tasks. In: The 30th EMBS Conference (EMBS 2008), pp. 1155–1158 (2008)
8. Lamberts, J., Van Den Broek, P.L.C., Bener, L., Van Egmond, J., Dirksen, R., Coenen, A.M.: Correlation dimension of the human Electroencephalogram corresponds with cognitive load. Neuropsychobiology 41, 149–153 (2000)
9. Stam, C.J.: Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. Clinical Neurophysiology 116, 2266–2301 (2005)
10. Natarajan, K., Acharya, R., Alias, F., Tiboleng, T., Puthusserypady, S.K.: Nonlinear analysis of EEG signals at different mental states. BioMedical Eng. Online 3, 1–11 (2004)
11. Nai-Jen, H., Palaniappan, R.: Classification of mental tasks using fixed and adaptive autoregressive models of EEG signals. In: The 2nd EMBC Conference on Neural Eng., pp. 633–636 (2005)
12. Zarjam, P., Epps, J., Lovell, N.H.: Characterising mental load in an arithmetic task using entropy-based features. In: The 11th ISSPA Conference (ISSPA 2012), pp. 245–250 (2012)
13. Zarjam, P., Epps, J., Lovell, N.H., Chen, F.: Characterization of memory load in an arithmetic task using non-linear analysis of EEG signals. To be Appeared in the EMBC 2012 (August 2012)
14. Rosso, O.A., Martin, M.T., Plastino, A.: Brain electrical activity analysis using wavelet-based informational tools. Physica A: Statistical Mechanics and its Applications 313, 587–608 (2002)
15. Rosso, O.A., Martin, M.T., Plastino, A.: Brain electrical activity analysis using wavelet-based informational tools (II): Tsallis non-extensivity and complexity measures. Physica A: Statistical Mechanics and its Applications 320, 497–511 (2003)
16. Martin, M.T., Plastino, A., Rosso, O.A.: Statistical complexity and disequilibrium. Physics Letters A 311, 126–132 (2003)
17. Zhukov, L., Weinstein, D., Johnson, C.: Independent component analysis for EEG source localization. Eng. in Medicine and Biology Magazine 19, 87–96 (2000)
18. Rosso, O.A., Martin, M.T., Figliola, A., Keller, K., Plastino, A.: EEG analysis using wavelet-based information tools. Journal of Neuroscience Methods 153, 163–182 (2006)
19. Stipacek, A., Grabner, R.H., Neuper, C., Fink, A., Neubauer, A.C.: Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. Neuroscience Letters 353, 193–196 (2003)
20. Harmony, T., Fernández, T., Silva, J., Bosch, J., Valdés, P., Fernández-Bouzas, A., Galán, L., Aubert, E., Rodríguez, D.: Do specific EEG frequencies indicate different processes during mental calculation? Neuroscience Letters 266, 25–28 (1999)
21. Shiner, J.S., Davison, M., Landsberg, P.T.: Simple measure for complexity. Physical Review E 59, 1459–1464 (1999)
22. Harmony, T., Fernández, T., Silva, J., Bernal, J., Díaz-Comas, L., Reyes, A., Marosi, E., Rodríguez, M., Rodríguez, M.: EEG delta activity: An indicator of attention to internal processing during performance of mental tasks. Int. J. Psych. 24, 161–171 (1996)

Attachment C

CHARACTERIZING MENTAL LOAD IN AN ARITHMETIC TASK USING ENTROPY-BASED FEATURES

Pega Zarjam [1,2] *, Julien Epps* [1,2]*, and Nigel H. Lovell* [3,1]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia, Eveleigh, NSW 2015, Australia
[3]Graduate School of Biomedical Eng. The University of NSW, Sydney, NSW 2052, Australia

ABSTRACT

We propose the use of entropy-based features; spectral and approximate entropies, of recorded EEG signals to characterize mental load when performing a cognitive task. It is demonstrated on a seven load-level task that the spectral entropy is a good discriminator of mental load level and decreases consistently in accordance with the increased load. The extracted approximate entropy quantifies the irregularity of the EEGs, indicating a decrease in irregularity as the load increases. We also perform EEG source estimation to choose a smaller subset of EEG channels which make the most contribution in the load level discrimination. We conclude that the entropy-based features are capable of measuring the imposed mental load from the selected channels in two brain regions. This may demonstrate that the brain behaves in a more regular or focused manner when dealing with higher task loads. The efficacy of entropy-based features across frequency sub-bands is also investigated.

1. INTRODUCTION

Measuring the amount of mental demand on the working memory when doing a cognitive process is of great importance for the prevention of decision-making errors, and the development of adaptive user interfaces [1]. This is to avoid mental overload and maintain efficiency and productivity in work performance, especially in critical/high mental load workplaces such as air traffic control, military operations, and so on.

A widely used brain monitoring technique for measuring cognitive workload is Electroencephalography (EEG), which offers high temporal resolution, ease of use, and a comparably low cost [2]. Finding features that are good discriminators of different workloads is another important key to successfully measuring and classifying the mental load.

Previously, a range of spectral features have been deployed for this purpose using EEG signals, including power spectral density (PSD), average power and maximum/log power spectra [3-5]. Entropy-based measures such as the wavelet packet entropy/entropy synchronization have been also used, but mainly in mental task classification [6, 7] or approximate entropy (ApEn) in pathology applications [8, 9].

For this study, we designed a cognitive task, more specifically an arithmetic task, with seven levels of difficulty to examine the performance of the entropy-based features for fine load level measurement and discrimination. To our knowledge the largest number of mental task load levels induced to date is five levels [10, 11], and our work advances this to seven. As further motivation, our earlier work with three levels on a reading task showed very promising results [12-14].

We also apply the concept of source localization to select a smaller number of EEG channels to inform the optimal channel selection. This concept has been used in medical applications previously to estimate the localization and distribution of electrical events in neural pathologies such as multifocal epilepsy and Alzheimer's [15-17], or anxious/depressive disorders [18].

This study was thus undertaken to examine the feasibility of applying entropy-based features to assess finer discrimination of mental load levels and to test the hypothesis that regularity/complexity of the recorded EEG signals changes as the task load varies. So far, the regularity or complexity of EEG signals in pathology/different mental tasks has been evaluated by nonlinear measures, like correlation dimension (CD), Hurst exponent (H), and ApEn [8, 19-21]. These measures have shown their effectiveness in understanding the complex dynamical behavior of the brain [20]. The related EEG channels and frequency bands for which reliable information may be extracted using entropy features for an EEG-based mental load measurement system were also investigated.

978-1-4673-0382-8/12/$31.00 ©2012 IEEE

2. MATERIALS

2.1. Experiment and Subjects

Six healthy male volunteers, 24-30 years of age, engaged in postgraduate study, participated in the experiment. We designed an addition task with seven levels of difficulty, starting from one digit addition (very low) to multi-digit addition (extremely difficult).

This addition task was displayed and controlled on a laptop PC with a viewing distance of 70 cm to the participant (subject). Each number was shown at the center of the screen in Arabic notation for three seconds. Subjects were asked to add the two presented numbers (shown sequentially), then were given two seconds (blank page) for retention followed by a multiple choice menu that presented the possible answers. The subjects were required to click on the correct answer using the mouse left button, using the minimum possible finger movement. There were a total of 42 addition problems, in seven difficulty levels, each level lasting for two minutes. The difficulty level was manipulated by varying the n-digit numbers used and carries required to calculate the addition. The task detail is shown in Table I.

To minimize any muscle movement artifact (EMG) during the recording, the participants were asked to avoid any unnecessary physical movements and their hand was placed in a fixed position, where they could still make finger movements in response to the correct answer on the mouse. Since the channels in the frontal lobes are susceptible to ocular artifact, participants were required to refrain from blinking as much as possible. The participants were given 30 second rests between each level, allowing them to relax, move or blink.

2.2. Data Acquisition

The participants' EEG signals were recorded using an Active Two acquisition system [13], at the ATP Laboratory of National ICT Australia in Sydney.

The experiment was conducted under controlled conditions in an electrically isolated laboratory, with a minimum distance of five meters from power sources to the experiment desk and under natural illumination. Each recording contained 32 EEG channels mounted in an elastic cap, according to the extended international 10 - 20 system. A linked earlobe reference was used and impedance was kept under $5k\Omega$. The EEG signals were passed through a band-pass filter with cut-off frequencies of $0.1 - 100\,Hz$ and were recorded at a $f_s = 256\,Hz$ sampling rate.

Each recording was judged by visual inspection to choose the epochs which contained minimal EMG artifact. As a result, 70 seconds (out of 90 seconds of each task level recording) for each subject was considered. However, the remaining portion of the recordings still included EOG and ECG artifacts.

Table I. The experimental presentation is shown here. The difficulty level was manipulated by varying the number of digit numbers used, and carries required to calculate the addition. In each task level, 6 additions were presented.

Task level	Number of digits	Example
Very low (L1)	1&2 digit numbers	45+2
Low (L2)	1&2 digit numbers with 1 carry	54+9
Medium (L3)	2 digit numbers with 1 carry	67+42
Medium-High (L4)	2 digit numbers with 2 carries	39+65
High (L5)	2&3 digit numbers with 1 carry	377+32
Very high (L6)	2&3 digit numbers with 2 carries	76+347
Extremely high (L7)	3 digit numbers with 3 carries	983+748

3. METHODOLOGY

Our methodology includes the EEG signal source localization using the minimum norm estimate algorithm, sub-band filtering by Discrete Wavelet Transform (DWT) and entropy-based feature extraction from the EEG signals. The detail is as follows:

3.1. EEG Source Localization

EEG source localization is a non-invasive signal processing technique that measures EEGs at various locations on the scalp to estimate the current sources within the brain. It has previously been used to estimate the localization and distribution of electrical events in brain disorders [15-17]. There are various algorithms for EEG source localization, among which cortical source imaging using a minimum norm estimate is one of the most common [22].

We performed this method using the eConnectome software developed at Minnesota University [23], to select the channels which make the most contribution in discriminating the imposed task load, out of 32 channels recorded for each subject.

3.2. Sub-Band Filtering

The DWT provides a time-scale representation of a given signal, generated by dilation and translation of a mother Wavelet.

We selected the Daubechies-4 mother Wavelet, which is localized and symmetric and has a smooth thresholding effect [24], to decompose the EEG signals of length T seconds, into five levels (scales),

corresponding to the EEG frequency bands, as shown in Table II.

Table II. EEG frequency bands corresponding to each Wavelet scale.

Wavelet scale	Component	Freq. range (Hz)	EEG freq. band
1	detail	64-128	Gamma
	approximate	0-64	
2	detail	32-64	
	approximate	0-32	
3	detail	16-32	Beta
	approximate	0-16	
4	detail	8-16	Alpha
	approximate	0-8	
5	detail	4-8	Theta
	approximate	0-4	Delta

3.3. Feature Extraction

We denoted here the EEG segments under study in a particular sub-band as $x[n]; n = 1,2, \dots, N$. Each segment had a length of T seconds. The EEG segments were analyzed in different frequency sub-bands as explained in the previous Section.

Two entropy-based features were extracted from each EEG segment, since entropy is considered to be a measure of EEG signal complexity, and could act as a potential feature. These features could also provide some information stored in the signal's probability distribution [25]. The two features are as follows:

(a) Spectral Entropy: Spectral entropy (SpEn) was used to measure how the original power was distributed in a particular frequency sub-band. It is given by [25]:

$$\text{SpEn} = -\frac{1}{\log N_f}\sum_{k=1}^{N_f} P_x[k] \ln P_x[k]$$

where N_f is the number of frequency bins used in the estimation of the PSD of the signal, $k = \frac{2N_f}{f_s} f$, and $P_f[k]$ is an estimate of the PDF in the k^{th} frequency bin; i.e. $\frac{(k-1)f_s}{2N_f} < f < \frac{k f_s}{2N_f}$.

The PDF is calculated by normalizing the PSD estimate with respect to the total spectral power in each frequency sub-band. N_f is the number of frequency bins in the PSD estimate.

According to the above equation, the spectral entropy attains its peak when all the frequency bins contain the same power, and it becomes smaller as the signal power tends to concentrate in a particular frequency bin.

(b) Approximate Entropy: ApEn is a non-linear entropy estimators, indicating regularity or predictability of a time series. Small ApEn indicates predictability or regularity in the signal.

The feature ApEn of a given EEG segment; $x[n], n = 1,2, \dots, N$ is calculated using the following algorithm [26]. The parameters m and r represent the embedding dimension and the vector comparison distance, respectively.

1) Form the following $N - m$ vectors:
$x_i = [x[i]\ x[i+1]\ \dots\ x[i+m-1]]$;
$i = 1, \dots, N - m + 1$

2) For a given i,
a) For $j = 1,2, \dots, N - m + 1; j \neq i$, calculate the distance $d[x_i, x_j]$ as:
$$d[x_i, x_j] = \max|x[k] - x[l]|;$$
$$k = i, i+1, \dots, i+m-1,$$
$$l = j, j+1, \dots, j+m-1$$
b) Find $C_i(m,r)$ as the number of X_i such that: $d[X_i, X_j] \leq r$
c) Next i.

3) Find: $\Phi(m,r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \log\left(\frac{C_i(m,r)}{N-m+1}\right)$

4) Repeat steps 1) to 3) for $m = m + 1$ and calculate $\Phi(m+1, r)$.

5) Calculate ApEn as:
$$\text{ApEn(m, r)} = \Phi(m,r) - \Phi(m+1,r).$$

The values of the parameters m and r are critical in determining the outcome [8].

ApEn has been shown to be robust to noise and also finite for stochastic, noisy deterministic and composite processes unlike its counterparts (e.g. Kolmogorov–Sinai entropy). Increasing values of ApEn relate to more irregularity or unpredictability [8].

4. EXPERIMENTAL RESULTS AND DISCUSSION

The source localization results showed that mainly the frontal and occipital regions of the brain in all the task load levels were the most influenced regions, across all six subjects. However, as the load level increased, not only wider areas of these regions were affected, but also they were affected more deeply (shown by values closer to "1" in Fig. 1). For illustration purposes, the source map of two load levels; the lowest (L1) and the most difficult levels (L7) for subject 1 are displayed in Fig. 1.

Thus, we only considered the EEG channels located in the frontal and occipital loops for further analysis, namely; the frontal left; channels Fp1, AF3, F7, F3, FC1, FC5 and the frontal right; channels FC6, FC2, F4, F8, AF4, Fp2, and the occipital lobe channels (channels PO3,

O1, Oz, O2, PO4). Then, the features defined in the previous Section were extracted from each EEG segment of $T = 5$ seconds of the above selected channels for different frequency sub-bands.

Fig. 2 displays the median of the extracted SpEn feature from the frontal channels in scale 5, across all subjects. As shown, the median of the SpEn is able to distinguish the seven task loads well across most of the EEG frontal channels in the delta frequency band, as it consistently exhibits a decreasing median with increasing task load (task difficulty).

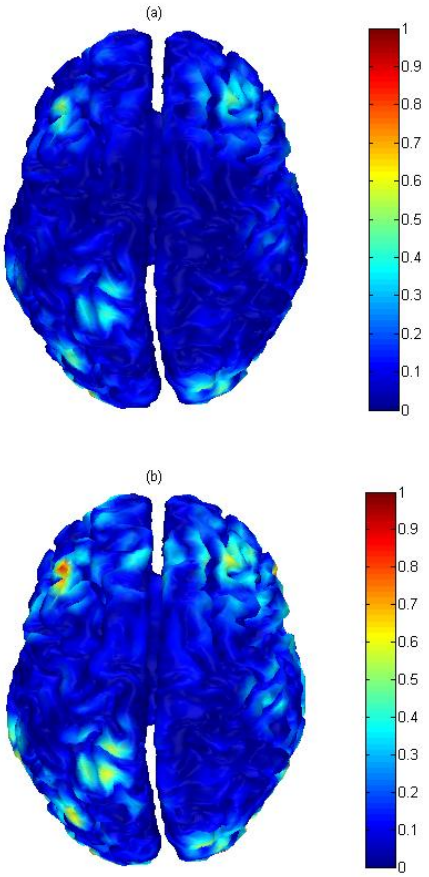


Fig. 1. The source maps of two load levels; (a) the lowest load (L1), and (b) the most difficult load (L7). As seen both load levels influence the similar regions more or less but the activation seems to increase as the load level increased.
Note that the scale has been normalized, so that "0" reflects no difference to the background colour and "1" reflects the maximum difference.

The results for other sub-bands are summarized in Table III. This table displays the channels which exhibit a regular trend (decreasing trend) as task load increases. This decrease indicates that the signal power tends to concentrate in a particular frequency bin, as the load increases.

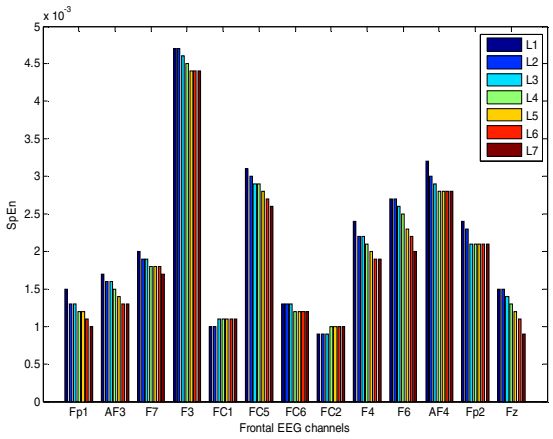


Fig. 2 Medians of the SpEn extracted from segmented EEG data in the delta band, from the frontal lobe across all subjects. In most of the frontal EEG channels, the associated median decreases as the task load increases. L1 represents the very low task level, and L7 the extremely difficult task level.

The ApEn was estimated here with $m = 2$ and $r = 0.2 *$ standard deviation (SD) of the EEG segment. These are the suggested values in [26] for studying EEG signals. This feature also exhibited a decreasing trend as the task load level increased across the channels under study. This decrease in the value of this feature shows that the signal's irregularity declines as the task load increases. In other words, the signals become more predictable as the task load or difficulty increases. Fig. 3 illustrates the median of the extracted ApEn from an occipital channel in scale 5, for subject 1. The results across all subjects for other sub-bands are summarized in Table III.

Investigating the features by sub-band, the delta sub-band exhibits more channels that consistently decrease for both extracted features, especially in the frontal loop. This is followed by fewer channels in the occipital loop for both features. The alpha sub-band is the second frequency band reflecting more channels for the SpEn in the occipital loop, than in the frontal loop. No channels were found useful for ApEn feature in this sub-band. The beta is the third sub-band reflecting few channels for the SpEn feature in both brain loops. The theta sub-band revealed no channels in the frontal loop for any feature, but a couple for the SpEn in the occipital loop.

In terms of the brain region investigation, the frontal loop revealed the highest number of channels contributing in the task load discrimination, for both extracted features. This is supported by similar findings that the increasing workload in the working memory is reflected by activity in the brain frontal region [10, 27].

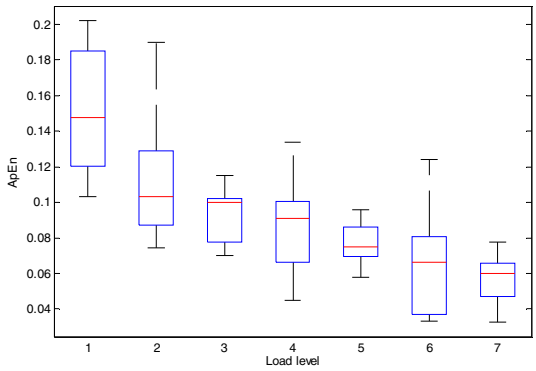


Fig. 3 Medians of the ApEn extracted from segmented EEG data in the delta band, from one occipital channel (PO3) of subject 1. On each box, the red mark is the median; the edges of the box are the 25th and the 75th percentiles.

Table III. Selected channels for each extracted feature showing a consistent decreasing trend with task load increase, in different frequency sub-bands, across all six subjects.

Brain region	Freq. sub-band	Feature	Channels
Frontal	Delta	SpEn	Fp1, AF3, F7, F3, FC5, F4, F8, Fz
		ApEn	Fp1, AF3, F7, F4,F8, AF4, Fp2, Fz
	Theta	SpEn	-
		ApEn	-
	Alpha	SpEn	F7, FC5, Fz
		ApEn	-
	Beta	SpEn	F3, FC5
		ApEn	-
Occipital	Delta	SpEn	PO3, O1, PO4
		ApEn	O1, Oz
	Theta	SpEn	PO4, O1
		ApEn	-
	Alpha	SpEn	PO3, O1, Oz, O2, PO4
		ApEn	-
	Beta	SpEn	PO3, O1, PO4
		ApEn	-

5. CONCLUSION

In this study, we proposed the use of two entropy-based features for characterizing mental load in an arithmetic task. The task was designed using seven finely-spaced levels to impose a large amount of varying mental workload on the subjects working memory. The source localization was applied to select the optimal channels among the 32 recorded channels which make the most contribution in the task load measurement. The results showed that across all task levels the frontal and occipital channels were affected the most when the task complexity was varied. As anticipated, with higher load levels these regions were influenced more deeply and widely.

The extracted features; namely SpEn and ApEn, were found to be successful in characterizing the task loads by showing a consistent decreasing value as the task load increased. Furthermore, the ApEn decline with the task load increase, could demonstrate the decreasing complexity or increasing regularity of the EEG signals. This may show that the brain behaves in a more regular or focused manner when performing more difficult tasks.

The frequency sub-band study revealed that the delta sub-band is the most significantly contributing frequency sub-band in the task load measurement, including more channels for both features.

In conclusion, a smaller number of channels in just two brain regions, in the delta sub-band could provide sufficient information for mental task load measurement in similar contexts using the entropy-based features.

However, this should be validated on a larger database and in more realistic environments. Future work includes collection of EEG signals with increased subject numbers, running different cognitive tasks with a focus on cognitive overload, using a classification method for discriminating the task loads, and finding the optimal m and r parameters for estimating the ApEn feature. Investigating the usefulness of the other nonlinear measures in mental load characterization is also another future objective.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge the volunteers for participating in the experiment. This research was supported by the Asian Office of Aerospace Research & Development, Grant No. FA2386-10-1-4029.

7. REFERENCES

[1] F. Paas, J. E. Tuovinen, H. Tabbers, *et al.*, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist,* vol. 38, pp. 63-71, 2003.

[2] O. T. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general " *Journal of Neural Eng.* , vol. 8, pp. 1-5, 2011.

[3] P. F. Diez, E. Laciar, V. Mut, *et al.*, "A comparative study of the performance of different spectral estimation methods for classification of mental tasks," in *the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (EMBS'08)*, pp. 1155-1158, 2008.

[4] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Research Reviews,* vol. 29, pp. 169-195, 1999.

[5] K. M. Spencer and J. Polich, "Poststimulus EEG spectral analysis and P300: Attention, task, and probability," *Psychophysiology,* vol. 36, pp. 220-232, 1999.

[6] Z. Li and M. Shen, "Classification of mental task EEG signals using wavelet packet entropy and SVM," in *the 8th International Conference on Electronic Measurement and Instruments, (ICEMI '07)* , pp. 3-906-3-909, 2007.

[7] C. J. Stam, A.-M. Van Cappellen Van Walsum and S. Micheloyannis, "Variability of EEG synchronization during a working memory task in healthy subjects," *International Journal of Psychophysiology,* vol. 46, pp. 53-66, 2002.

[8] D. Abásolo, R. Hornero, P. Espino*, et al.*, "Analysis of regularity in the EEG background activity of Alzheimer's disease patients with approximate entropy," *Clinical Neurophysiology,* vol. 116, pp. 1826-1834, 2005.

[9] O. Hasan, "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy," *Expert Systems with Applications,* vol. 36, pp. 2027-2036, 2009.

[10] A. Stipacek, R. H. Grabner, C. Neuper*, et al.*, "Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load," *Neuroscience Letters,* vol. 353, pp. 193-196, 2003.

[11] C. Berka, D. J. Levendowski, M. N. Lumicao*, et al.*, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, Space, and Environmental Medicine,* vol. 78, pp. B231-B244, 2007.

[12] P. Zarjam, J. Epps and F. Chen, "Characterizing working memory load using EEG delta activity," presented at the 19th European Signal Processing Conference (Eusipco'11), Barcelona, Spain, 2011.

[13] P. Zarjam, J. Epps and F. Chen, "Spectral EEG features for evaluating cognitive load," presented at the Proc. of the 33rd Annual International Conference of the IEEE EMBS (EMBS'11), Massachusetts USA, 2011.

[14] P. Zarjam, J. Epps and F. Chen, "Evaluation of working memory load using EEG signals," presented at the Second Asia Pacific Signal Processing Conference (APSIPA'10), Singapore, 2010.

[15] L. Zhukov, D. Weinstein and C. Johnson, "Independent component analysis for EEG source localization," *Engineering in Medicine and Biology Magazine, IEEE,* vol. 19, pp. 87-96, 2000.

[16] C. Phillips, M. D. Rugg and K. J. Friston, "Systematic regularization of linear inverse solutions of the EEG source localization problem," *NeuroImage,* vol. 17, pp. 287-301, 2002.

[17] T. Dierks, V. Jelic, R. D. Pascual-Marqui*, et al.*, "Spatial pattern of cerebral glucose metabolism (PET) correlates with localization of intracerebral EEG-generators in Alzheimer's disease," *Clinical Neurophysiology,* vol. 111, pp. 1817-1824, 2000.

[18] M. Ruchsow, B. Herrnberger, P. Beschoner*, et al.*, "Error processing in major depressive disorder: Evidence from event-related potentials," *Journal of Psychiatric Research,* vol. 40, pp. 37-46, 2006.

[19] K. Natarajan, R. Acharya, F. Alias*, et al.*, "Nonlinear analysis of EEG signals at different mental states," *BioMedical Enineering Online,* vol. 3, pp. 1-11, 2004.

[20] C. J. Stam, "Brain dynamics in theta and alpha frequency bands and working memory performance in humans," *Neuroscience Letters,* vol. 286, pp. 115-118, 2000.

[21] J. Jeong, "EEG dynamics in patients with Alzheimer's disease," *Clinical Neurophysiology,* vol. 115, pp. 1490-1505, 2004.

[22] R. Grech, T. Cassar and J. Muscat, "Review on solving the inverse problem in EEG source analysis," *Journal of NeuroEng. and Rehabilitation,* vol. 5, pp. 1-33.

[23] B. He, Y. Daia and L. Astolfic, "eConnectome: A MATLAB toolbox for mapping and imaging of brain functional connectivity.," ed: University of Minnesota, 2011.

[24] M. Akay, I. E. I. Medicine and B. Society, *Time frequency and wavelets in biomedical signal processing*: IEEE Press, 1998.

[25] B. R. Greene, S. Faul, W. P. Marnane*, et al.*, "A comparison of quantitative EEG features for neonatal seizure detection," *Clinical Neurophysiology,* vol. 119, pp. 1248-1261, 2008.

[26] S. M. Pincus, "Assessing serial irregularity and its implications for health," *Annals of the New York Academy of Sciences,* vol. 954, pp. 245-267, 2001.

[27] T. Harmony, T. Fernández, J. Silva*, et al.*, "Do specific EEG frequencies indicate different processes during mental calculation?," *Neuroscience Letters,* vol. 266, pp. 25-28, 1999.

34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Characterization of Memory Load in an Arithmetic Task using Non-Linear Analysis of EEG Signals

Pega Zarjam, *Student Member, IEEE,* Julien Epps, *Member, IEEE,* Nigel H. Lovell, *Fellow, IEEE* and Fang Chen

***Abstract*— In this paper, we investigate non-linear analysis of electroencephalogram (EEG) signals to examine changes in working memory load during the performance of a cognitive task with varying difficulty levels. EEG signals were recorded during an arithmetic task while the induced load was varying in seven levels from very easy to extremely difficult. The EEG signals were analyzed using three different non-linear/dynamic measures; namely: correlation dimension, Hurst exponent and approximate entropy. Experimental results show that the values of the measures extracted from the delta frequency band of signals acquired from the frontal and occipital lobes of the brain vary in accordance with the task difficulty level induced. The values of the correlation dimension increased as the task difficulty increased, showing a rise in complexity of the EEG signals, while the values of the Hurst exponent and approximate entropy decreased as task difficulty increased, indicating more regularity and predictability in the signals.**

I. Introduction

A RELIABLE and noninvasive measurement of working memory load that can be made continuously while performing a cognitive task would be very helpful for assessing cognitive function, crucial for the prevention of decision-making errors, and the development of adaptive user interfaces [1]. Such a measurement could help to maintain the efficiency and productivity in task completion, work performance, and to avoid cognitive overload [1], especially in critical/high mental load workplaces such as air traffic control, military operations, and fire/rescue commands.

Electroencephalography (EEG) is a noninvasive neuro-imaging technique widely used for measuring cognitive workload, which offers high temporal resolution, ease of use, and a comparably low cost [2]. EEG contains useful information about various physiological states of the brain and can be very efficient for understanding the complex dynamical behavior of the brain, if interpreted correctly [3].

This work was supported by the Asian Office of Aerospace Research & Development, Grant No. FA2386-10-1-4029.

P. Zarjam is with the School of EE&T, University of New South Wales, Sydney, Australia. She is also with ATP Research Laboratory, National ICT Australia (phone: +61 2 9385 4803; fax: +61 2 9385 5993; email: p.zarjam@student.unsw.edu.au).

J. Epps is with the School of EE&T, University of New South Wales, Sydney, Australia. He is also with ATP Research Laboratory, National ICT Australia (email: j.epps@unsw.edu.au).

Nigel Lovell is with the Graduate School of Biomedical Eng. University of New South Wales, Sydney, Australia (email: n.lovell@unsw.edu.au).

F. Chen is with ATP Research Laboratory, National ICT Australia (email: fang.chen@nicta.com.au).

Previously, a range of methods have been applied for measuring and classifying the memory load using EEG signal. These methods have used features such as power spectral density (PSD) or the averaged power and maximum/log power spectra [4-6], sub-band entropy [7-8], and autoregressive model [9]. The application of non-linear methods in classifying mental tasks is more recent, and measures like correlation dimension (CD) [10-12], Hurst exponent (HE), approximate entropy (ApEn) and largest Lyapunov exponent (LLE) [13-14] have been used to measure the complexity/irregularity of the underlying brain dynamics during the performance of some cognitive tasks compared with the rest condition. In [13], it was demonstrated that the CD and ApEn/HE values decrease/increase when the participants are subject to sound or reflexologic stimulation compared with the normal state, showing a lesser degree of cognitive activity. Stated differently, in these studies the brain activity states; such as normal/rest and stimulated have been differentiated [10, 13-14]. But to date, these measures have not been investigated in the analysis of the varying working memory load and the question whether these approaches could provide some information on the brain dynamics/behavior when performing a cognitive task with varying difficulty levels has not been addressed.

For this study, we designed a cognitive task, more specifically an arithmetic task with seven levels of difficulty. To our knowledge the largest number of mental task load levels reported to date is five levels [15-16]. Our earlier work with three levels on a reading task also showed very promising results in characterizing the memory load using linear features [17-18].

We hypothesize that non-linear measures change continuously according to the varying difficulty levels of the cognitive task induced and therefore they can be used to quantify changes in memory loads during the performance of a cognitive task.

II. Non-Linear Measures Background

In this study, we analyze the EEG signals during the performance of an arithmetic task using CD, HE, and ApEn. The measures are briefly explained below. Full details of their computation and the selection of their parameters can be found in [11, 13, 19].

Correlation dimension (CD): this is a measure of the complexity of a time series. For a given EEG segment; N, CD is a function of two parameters;

d r, which represent the embedding dimension and radial space around each reference point, respectively. The CD is calculated using [11]:

_____ (1)

where is a function showing the probability that two arbitrary points of in an m-dimensional space on the orbit are closer together than Larger values of CD indicate more complexity in the signal.

Approximate entropy (ApEn): this is a non-linear entropy estimator showing regularity or predictability of a given time series. ApEn of a given $x[n]$; is calculated using the following formula [19]:

(2)

____ ____ (3)

Here,)= (number of *X(j)* such that r. is the max distance between two given vectors of *X(i), X(j).* A parameter *m*

$+1$. Practically, ApEn quantifies the likelihood of vectors that remain close (within *r*) on the next incremental comparison [20]. Larger values of ApEn indicate unpredictability or irregularity in the signal.

Hurst exponent (HE): it is a measure of self-similarity and long-term dependence and its degrees in a time-series. It is defined by [13]: –

where $T=N*f_s$ is the duration of the sample data and – the corresponding value of the rescaled range. If the time-series covers more distance than a random walk. Larger values of HE represent increase of randomness in the signal.

III. METHODS

A. Participants and Experiment Settings

We studied six male participants, between the ages of 24-30 years, engaged in postgraduate study. They were right-handed and had normal or corrected to normal eyesight and gave written informed consent, in accordance with human research ethics guidelines. We designed an addition task with seven levels of difficulty, starting from one digit addition (very low) to multi-digit addition (extremely difficult).

The task was displayed and controlled on a laptop PC with a viewing distance of 70 cm to the participant (subject). Each number was shown at the center of the screen in Arabic notation for three seconds. Subjects were asked to add the two presented numbers (shown sequentially), then were given two seconds (blank page) for retention followed by a multiple choice menu that presented the possible answers. The subjects were required to click on the correct answer using the mouse left button, with the minimum possible finger movement. There were 42 addition problems in total, across seven difficulty levels (6 per level), with each level lasting for two minutes. The difficulty level was manipulated by varying the n-digit numbers used and carries required to calculate the addition. The task detail is shown in Table I.

TABLE I.
TASK DIFFICULTY LEVEL DETAILS.

Task level	Number of digits	Example
Very low (L1)	1&2 digit numbers	45+2
Low (L2)	1&2 digit numbers with 1 carry	54+9
Medium (L3)	2 digit numbers with 1 carry	67+42
Medium-High (L4)	2 digit numbers with 2 carries	39+65
High (L5)	2&3 digit numbers with 1 carry	377+32
Very high (L6)	2&3 digit numbers with 2 carries	76+347
Extremely high (L7)	3 digit numbers with 3 carries	983+748

The participants were asked to avoid any unnecessary physical movements to minimize the chance of muscle movement artifact (EMG) during the recording. Their hand was also placed in a fixed position, where they could still make finger movements in response to the correct answer on the mouse. Since the channels in the frontal lobes are sensitive to ocular artifact, participants were required to refrain from blinking as much as possible. The participants were given 30 second rests between each level, allowing them to relax, move or blink.

B. EEG Recording

The EEG signals were recorded from 32 channels mounted in an elastic cap, according to the extended international 10 - 20 system using an Active Two acquisition system. The experiment was conducted under controlled conditions in an electrically isolated laboratory, with a minimum distance of five meters from power sources to the experiment desk and under natural illumination. The EEG signals were passed through a band-pass filter with cut-off frequencies of Hz and were recorded at a $f_s = 256\,Hz$ sampling rate. Each recording was visually inspected to choose the epochs which contained minimal EMG artifact. As a result, 70 seconds (out of 90 seconds of each task level recording) for each subject was considered. However, the remaining portion of the recordings still included EOG and ECG artifacts.

IV. ANALYSIS

A. EEG Source Localization

We used EEG source localization to estimate the localization and distribution of electrical events to select discriminatory channels, as in our previous work [21].

B. Sub-Band Filtering

We decomposed the EEG signals using the Discrete Wavelet Transform (DWT) into five levels (scales), according to the EEG frequency bands (0-4Hz delta, 4-8 Hz theta, 8-12 Hz alpha, 12-30 Hz beta, 30-100 Hz gamma). The selected mother wavelet was the Daubechies-4, which is localized and symmetric and has a smooth thresholding effect.

C. Non-Linear Measure Application

The EEG segments in a particular sub-band were denoted as

$, N$ with the length of $T = 5$ seconds. Three non-linear measures; i.e. CD, ApEn, and HE were extracted from each EEG segment in different frequency sub-bands for each subject.

V. Results

The source localization results showed that mainly the frontal and occipital regions of the brain were the most influenced regions, in all the task load levels across all six subjects. As the load level increased, not only were wider areas of these regions were affected, but also they were affected more deeply (shown by values closer to "1" in Fig. 1). The source maps of two load levels, the lowest (L1) and the most difficult levels (L7) for subject 1, are shown in Fig. 1. Therefore, for further analysis only EEG channels positioned in the frontal and occipital lobes were taken into account (i.e. the frontal channels Fp1, AF3, F7, F3, FC1, FC5 FC6, FC2, F4, F8, AF4, Fp2 and the occipital channels PO3, O1, Oz, O2, PO4).

Fig. 2(a) shows the medians of the extracted CD measure from a frontal channel for subject 1 in the delta frequency band. As seen, the median of the CD increases regularly as the task load increases. Fig. 2(b) displays the median of the extracted ApEn measure from the same channel, subject and frequency band. Here, the median of the ApEn decreases consistently as the task load increases. The extracted HE values showed a similar trend to the ApEn. Therefore, their values tended to decline as the load level increases.

The results for the selected channels across all the six subjects in different sub-bands are summarized in Table II. The study of these measures by frequency sub-band indicated that the delta sub-band exhibited more channels that consistently vary with the load level induced. In terms of the brain regions investigated, the frontal lobe also showed the highest number of channels contributed to the load level distinction.

Due to the importance of the non-linear parameters' values in determining the outcome, we also examined their different values to find the optima for the purpose of memory load characterization in this study. Thus, we calculated the CD for $1 < m < 10$ and $r > 10$. According to the results, the higher the dimension m, the more distinct the load levels were. But varying parameter r did not affect the results much. For the ApEn measure, we varied $0.$ td and $m = 2$ or 3. The results showed that the lower the r value (closer to $0.1 * std$), the better the load levels were distanced but the choice of embedding dimension of 2 or 3 did not make any significant change.

We also used a Kruskal-Wallis test to statistically measure the effectiveness of the measures in distinguishing seven load levels. The channels which revealed a small p-value ($p < 0.01$) for each extracted measure, across six subjects are shown in bold in Table II.

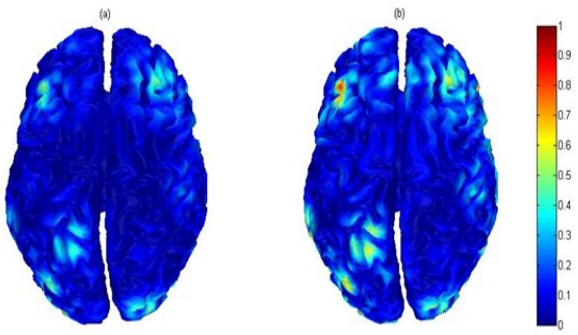


Fig. 1. The source maps of two load levels for subject 1; (a) the lowest load (L1), and (b) the most difficult load (L7). Both load levels influence the similar regions more or less but the degree of activation increased as the load level increased.

VI. Discussion

In this study, we investigated the use of three non-linear measures for characterizing memory load in an arithmetic task with seven levels of difficulty. The source localization results assisted us in focusing on the brain regions/channels of interest which were the most influenced by the task load, namely the frontal and occipital lobes. When the more difficult task load was induced these regions were affected more deeply and widely. This is in line with previous findings that the increasing workload is reflected by activity mostly in the frontal lobe of the brain [15, 22].

The extracted non-linear measures from the selected channels were found to be successful in task load discrimination and representing the functional dynamics of the brain when performing a task with different difficulty levels. The CD values tended to increase as the task load increased; indicating the brain activity dimension/complexity increases with the increase of cognitive activity load. This can be supported by previous mental task studies showing lower dimension when the brain goes to a passive state or a state of relaxation [10, 13]. A decreased value of ApEn with increased task load implies higher predictability and less irregularity in the brain activity. The decline in HE values as the task load increased demonstrates that random behavior of the signal decreases as the task load increases. The last two measures may indicate the brain behaves in a more regular and focused manner when performing more difficult tasks.

The frequency sub-band analysis showed that the delta is the most contributing sub-band, including more channels for the three measures in the memory load characterization. This was statistically confirmed by low p-values.

As future work, this method should be validated on a larger database and in more realistic environments. This includes collection of EEG signals with increased subject numbers, running different cognitive tasks with a focus on cognitive overload, using a classification method for discriminating the task loads.

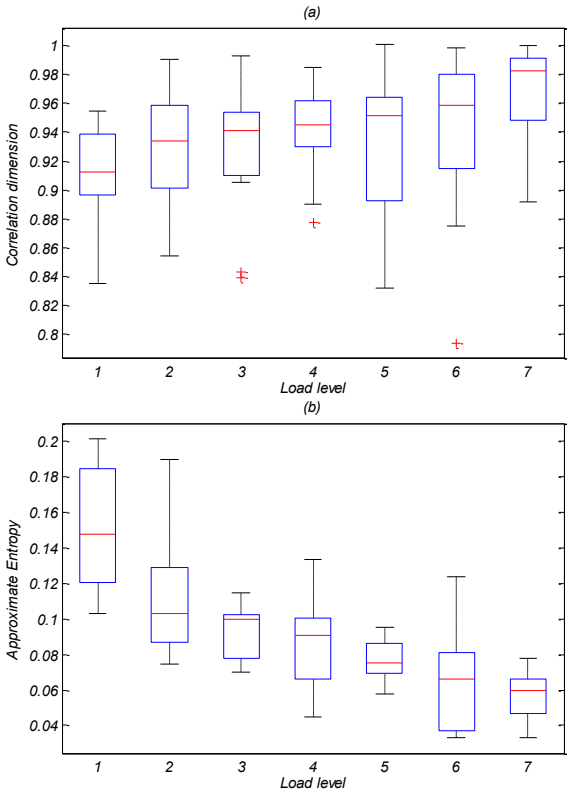


Fig. 2. (a) Medians of the CD ($m = 10, r = 20$) extracted from segmented EEG data in the delta band from a frontal channel (Fp1) of subject 1. (b) Medians of the ApEn ($m = 2, r = 0.2 * \text{std}$) extracted for the same channel, freq. band, and subject. On each box, the red mark is the median; the edges of the box are the 25th and the 75th percentiles.

TABLE II.
SELECTED CHANNELS FOR EACH EXTRACTED NON-LINEAR MEASURE WHOSE MEDIAN SHOWED A CONSISTENT TREND ACCORDING TO TASK LOAD VARIATION, IN DIFFERENT FREQUENCY SUB-BANDS, ACROSS ALL SIX SUBJECTS. CHANNELS IN BOLD DENOTE CASES WHERE A KRUSKAL-WALLIS TEST GAVE P<0.01 FOR ALL SIX SUBJECTS.

Freq. sub-band	Measure	Region of the brain/Channels
Delta	CD	Frontal: **Fp1**, AF3, FC1, F3, FC5, FC2, **F4**, **F8,** AF4, **Fp2** - Occipital: O1, **Oz**
	ApEn	Frontal: Fp1, **AF3**, **F7**, **F3**, FC5, **FC6**, FC2, F4, **F8**, **AF4**, Fp2 - Occipital: **O1**, Oz, **PO4**
	HE	Frontal: Fp1, AF3, FC1, **FC6**, F4, **F8**, AF4, Fp2 - Occipital: O1, **PO4**
Theta	CD	Frontal: **Fp1**, AF3, FC6, FC2 - Occipital: O1,Oz
	ApEn	Frontal: Fp1, AF3, FC1, F3, F8 - Occipital: O1, O2, PO4
	HE	Frontal: Fp1, AF3, FC1, FC2, F8 - Occipital: PO3,Oz, O2, PO4
Alpha	CD	-
	ApEn	Frontal: Fp1, AF3, F7, FC5, AF4, Fp2 - Occipital: PO3, O2
	HE	Frontal: FC1, FC6, F8 - Occipital: **Oz**
Beta	CD	-
	ApEn	Frontal: Fp1, AF3, FC6, Fp2 - Occipital: O1, Oz
	HE	Frontal: **Fp1**, **AF3**, FC2, Fp2 - Occipital: PO3

REFERENCES

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and *et al.* , "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist,* vol. 38, pp. 63-71, 2003.

[2] O. T. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general " *Journal of Neural Eng.,* vol. 8, pp. 1-5, 2011.

[3] O. Hasan, "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy," *Expert Systems with Applications,* vol. 36, pp. 2027-2036, 2009.

[4] P. F. Diez, E. Laciar, V. Mut, and *et al.*, "A comparative study of the performance of different spectral estimation methods for classification of mental tasks," in *the 30th EMBS Conference,* pp. 1155-1158, 2008.

[5] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Research Reviews,* vol. 29, pp. 169-195, 1999.

[6] K. M. Spencer and J. Polich, "Poststimulus EEG spectral analysis and P300: Attention, task, and probability," *Psychophysiology,* vol. 36, pp. 220-232, 1999.

[7] Z. Li and M. Shen, "Classification of mental task EEG signals using wavelet packet entropy and SVM," in *the 8th ICEMI Conference*, pp. 3-906-3-909, 2007.

[8] C. J. Stam, A.-M. van Cappellen van Walsum, and S. Micheloyannis, "Variability of EEG synchronization during a working memory task in healthy subjects," *Int. Journal of Psychophysiology,* vol. 46, pp. 53-66, 2002.

[9] H. Nai-Jen and R. Palaniappan, "Classification of mental tasks using fixed and adaptive autoregressive models of EEG signals," in *the 2nd Neural Engineering Conference,* pp. 633-636, 2005.

[10] J. Lamberts, P. L. C. van den Broek, L. Bener, and *et al.*, "Correlation dimension of the human Electroencephalogram corresponds with cognitive load," *Neuropsychobiology,* vol. 41, pp. 149-153, 2000.

[11] C. J. Stam, "Brain dynamics in theta and alpha frequency bands and working memory performance in humans," *Neuroscience Letters,* vol. 286, pp. 115-118, 2000.

[12] C. J. Stam, "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field," *Clinical Neurophysiology*, vol. 116, pp. 2266-2301, 2005.

[13] K. Natarajan, R. Acharya, F. Alias, and *et al.*, "Nonlinear analysis of EEG signals at different mental states," *BioMedical Eng. Online*, vol. 3, pp. 1-11, 2004.

[14] N. Kannathal, U. R. Acharya, C. M. Lim, and *et al.*, "Characterization of EEG—A comparative study," *Computer Methods and Programs in Biomedicine,* vol. 80, pp. 17-23, 2005.

[15] A. Stipacek, R. H. Grabner, C. Neuper, and *et al.*, "Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load," *Neuroscience Letters,* vol. 353, pp. 193-196, 2003.

[16] C. Berka, D. J. Levendowski, M. N. Lumicao, and *et al.* "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, Space, and Environmental Medicine,* vol. 78, pp. B231-B244, 2007.

[17] P. Zarjam, J. Epps, and F. Chen, "Spectral EEG features for evaluating cognitive load," in *the 33rd EMBS Conference*, pp. 3841-3844, 2011.

[18] P. Zarjam, J. Epps, and F. Chen, "Characterizing working memory load using EEG delta activity," in *the 19th Eusipco Conference*, pp. 1554-1558, 2011.

[19] D. Abásolo, R. Hornero, P. Espino, and *et al.*, "Analysis of regularity in the EEG background activity of Alzheimer's disease patients with approximate entropy," *Clinical Neurophysiology*, vol. 116, pp. 1826-1834, 2005.

[20] S. Pincus, "Approximate entropy as a complexity measure," Chaos (Woodbury, N.Y.), vol. 5, pp. 110-117, 1995.

[21] P. Zarjam, J. Epps, and N. Lovell, "Characterizing mental load in an arithmetic task using entropy-based features," to be presented at *the 11th ISSPA Conference*, July 2012.

[22] T. Harmony, T. Fernández, J. Silva, and *et al.,* "Do specific EEG frequencies indicate different processes during mental calculation?," *Neuroscience Letters,* vol. 266, pp. 25-28, 1999.

COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2012) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Automatic classification of eye activity for cognitive load measurement with emotion interference

Siyuan Chen[a,b,*], *Julien Epps*[a,b]

[a] *The School of Electrical Engineering and Telecommunications, The University of New South Wales, Kensington, NSW 2052, Australia*
[b] *National ICT Australia (NICTA), Level 5, 13 Garden Street, Eveleigh, NSW 2015, Australia*

ARTICLE INFO

Article history:
Received 20 March 2012
Received in revised form
27 October 2012
Accepted 31 October 2012

Keywords:
Cognitive load
Physiological measures
Eye activity
Pupil
Blink
Fixation
Saccade
Eye movement

ABSTRACT

Measuring cognitive load changes can contribute to better treatment of patients, can help design effective strategies to reduce medical errors among clinicians and can facilitate user evaluation of health care information systems. This paper proposes an eye-based automatic cognitive load measurement (CLM) system toward realizing these prospects. Three types of eye activity are investigated: pupillary response, blink and eye movement (fixation and saccade). Eye activity features are investigated in the presence of emotion interference, which is a source of undesirable variability, to determine the susceptibility of CLM systems to other factors. Results from an experiment combining arithmetic-based tasks and affective image stimuli demonstrate that arousal effects are dominated by cognitive load during task execution. To minimize the arousal effect on CLM, the choice of segments for eye-based features is examined. We then propose a feature set and classify three levels of cognitive load. The performance of cognitive load level prediction was found to be close to that of a reaction time measure, showing the feasibility of eye activity features for near-real time CLM.

1. Introduction

Interest in including cognitive technology in clinical practice has seen an increase in recent years. Common applications are the use of cognitive tests to assess the deficit when impairments occur in central nervous system neuropathology [40], for example, head injury [1], Schizophrenia [2], long-term alcohol abuse [3], Alzheimer disease and related disorders [4], to name a few. Moreover, cognitive assessment can also be of benefit in screening discharge patients [5] and in construction of individualized rehabilitation strategies [2], since cognitive skills are associated with daily living and social activities. As Spaulding et al. [2] have suggested, "a cognitive technology can be perfected that would contribute significantly to diagnosis, treatment and rehabilitation planning, evaluation of patients' response to treatment, and the design of future treatment modalities". Although the specification of function to be measured is different, evidence shows those aforementioned diseases or disorders are associated with memory capability [2–4]. Since cognitive load occurs as a result of the limited working memory available during a task [35], measuring cognitive load on patients in the cognitive tests can offer insights for patient treatments. For example, high cognitive load and short stimulus duration were found to create a critical performance distinction for schizophrenic patients [36].

Other applications include reducing medical errors due to high memory load on clinicians in the context of emergency

* *Corresponding author at*: The School of Electrical Engineering and Telecommunications, The University of New South Wales, Kensington, NSW 2052, Australia. Tel.: +61 403282498.
E-mail address: siyuan.chen@unsw.edu.au (S. Chen).
0169-2607/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2012.10.021

department. Studies have showed that the interruptions (cause information loss) and multitasking induce high cognitive load that contributes to medical errors [37]. Solutions proposed include using electronic tools to support adaptive process [37] on site and providing effective training [38] beforehand to reduce the cognitive load in work place.

Another focus is on evaluation of clinical information systems. Approaches are based on usability engineering and cognitive task analysis to ensure low cognitive load involved in use of such systems while users are carrying out tasks [39].

The development of an automatic cognitive load measurement (CLM) system is thus motivated by assessing user (patient) dynamic cognitive load, using psychophysiological and behavioral signals. Conventional methods for CLM, in particular subjective assessment, reaction time and performance (accuracy) cannot provide satisfactory results in all situations as they rely on overt responses without adequate temporal sensitivity, which they assume that users (patients) are willing to provide [6].

One fundamental problem that has limited the use of psychophysiological and behavioral signals for CLM to date is the presence of artifacts due to other mental resource demands [28]. Task-focused mental activity is not the only possible source of variation manifested in psychophysiological and behavior signals. For example, speech, heart rate variability, GSR and respiration are reported as effective features not only in CLM but also in emotion recognition and stress detection (e.g. [14,15]). For affective data, emotion is often elicited by stimuli with the task performance as a function of emotional state or as appraisals of a situation [16]. When collecting cognitive load data, task difficulty is carefully controlled with neutral (i.e. non-emotive) stimuli and emotional stimuli are avoided. An interesting question is what would happen to the psychophysiological and behavioral signals when users (patients) are performing a cognitive task and are subject to concurrent emotional stimuli. Such a question has important considerations in practice, where emotions and cognitive load cannot be expected to occur in isolation as they often do in the research laboratory – how should cognitive load classification systems be built that are robust to such types of variability?

The work in this paper is novel in (i) assessing eye pattern changes during tasks with emotional stimuli, with a view to validating eye activity-based CLM; (ii) recognizing the eye activity patterns for five levels of induced cognitive load, there by going beyond simply distinguishing low and high cognitive load levels; (iii) determining the eye feature dependence on arousal factors and the appropriate measurement timing for reliable load level estimation during task execution with interference from other sources.

2. Eye activity background and related work

2.1. *Advantage of using eye activity for CLM*

Four arguments are forwarded in favor of using eye activity patterns for CLM: (i) eye activity contains three classes of eye information, but still uses one sensor for data collection. Pupil dilation is a physiological signal whose changes are due to autonomic nervous system activity in the peripheral nervous system. Eye blink is a behavioral signal [17] (some papers also call it a psychophysiological response [19]) controlled by the central nervous system (CNS). Fixation and saccade are encoded by neural signals from cortical and subcortical systems. The different mechanisms could measure various underlying processes responsible for different aspects of cognitive activity. (ii) Eye activity is more ubiquitous than other modalities: we are free to use our eyes everywhere and anytime. (iii) Pupillary response and eye blink have been shown to correlate with both visual and aural cognitive tasks [9,17], thus can be applied in broad scenarios. (iv) Eye activity data collection is less intrusive than other physiological signal data collection. For example, eye tracking technology has been demonstrated to follow eye activity remotely [9].

2.2. *Pupillary response*

The basic function of pupil diameter change is to protect the retina (the light reflex) and also to respond to a shift in fixation from far to near objects (the near reflex). Changes that reflect variations in cognitive activities are relatively small compared with the changes due to light reflex and near reflex. In addition, the light reflex results in a relatively rapid pupillary response [20]. Therefore, if objects have nearly constant depth in the user's (patient's) visual field, we can consider the task-evoked pupillary response to comprise the low frequency components in the pupillary response spectrum.

Over a few decades of research on pupillary response, researchers still do not agree whether the pupil is a measure of emotional arousal or mental effort. Empirical studies found that pupil size increases as participants are exposed to more arousing images and sounds, regardless of valence [21,22]. Early research [23] on arousal and cognition attempted to manipulate some arousal factors while controlling the cognitive demands of tasks. They concluded that cognitive factors have a higher priority than the arousal factors in affecting pupil dilation. The arousal effect in pupillary response was not observed in sentence listening and addition tasks but only in the low cognitive load task, listening to countdown numbers [23]. However, in that experiment, tasks were controlled in auditory presentation and arousal levels were manipulated by the proximity of the stimulus (a word) to the subject of the sentences, reward or threat of electrical shock. The effect of auditory-induced emotion might be transient and not be as strong as in visual presentation, and the auditory based cognitive load might be higher compared with visual tasks, therefore we used affective images to induce controlled emotional effects.

2.3. *Eye blink*

Eye blinks occur only two to four times per minute for functional purposes [17]. There are other, non-functional types: reflexive blink (a protective response, e.g. to a puff of air), voluntary blink (a purposeful response depends on one's will) and endogenous blink (unconsciously occurs). The majority of eye blink behaviors are endogenous blinks, which are centrally controlled and have a link to cognition [17,20], therefore this type of blink is used for CLM. During a task-centered scenario, voluntary blinks can be avoided by

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

informing participants to behave naturally, while for reflexive blinks, selective analysis windows must be employed to include only important task-relevant stimuli, which may also reduce the number of functional blinks.

One finding is that blinking tends to be avoided to maximize stimulus perception during high-attention tasks, and blink occurrence is reduced with increasing information content and task demands [17]. Another explanation treats eye blink as a relief mechanism. In states of thinking nothing and task completion, eye blinks occur rarely because the 'mental tension' is relieved in the internal channel of solving problems. When the 'mental tension' cannot find an internal or external outlet, eye blink rate is increased [24]. Both claims can adequately explain most blink behaviors. Therefore we assume that the blink rate will increase in mental tasks with constant information content when the task difficulty increases, and try to ensure that the time window for blink feature extraction is long enough to allow measurable changes during task execution.

2.4. *Eye movement*

Fixation and saccade can be separated and labeled by automatic identification algorithms using the vertical and horizontal position of the pupil [25]. Fixation can be seen as a stationary state over regions of interest and is usually defined as the pause time above 100 ms within a range of 1° of visual angle. There is a general consensus that visual perception and cognitive processing occur during fixation. Saccades are rapid eye movements from one position to another, during which little visual processing occurs. Fixation and saccade occur in turn when eyes are viewing a scene [26,27].

Fixation duration is often used as a metric reflecting the difficulty of information extraction in the region of interest, while fixation rate indicates the degree of importance of the element. Another useful feature is saccade amplitude, which implies the difficulty of catching precise target positions [27].

2.5. *Approaches to eye-based cognitive load measurement*

Previous research on measuring cognitive load through eye tracking has often employed offline statistical analysis and focuses on the measurable magnitude of pupil dilation under carefully controlled cognitive task contexts. Studies show that pupil diameter can be used to identify group differences, for example, in intelligence [29], age [30] and visual and auditory presentations [9]. A recent work [42] used pupil diameter and fixation duration to assess workload in a simulated anesthetic care environment with the aim of improving safety in clinical practice. In the study, pupillary responses were observed in different stages of inducing general anesthesia at a high-fidelity simulator as the workload was increased by a critical incident. Although previous studies found the link between pupillary response and cognitive load, the influence of arousal during CLM, as another major source of mental activity affecting pupil dilation, has yet to be investigated. Valence can be another factor that affects cognitive activity. A study used pupillary response to validate emotion state with three valences, positive, neural and negative, during tasks and during tasks with an emotional avatar [44]. However, in this study, we are interested in variations in eye features with both arousal and valence induced by rated stimuli, many levels of task difficulty, and in examining whether they are a function of task difficulty. There are intensive studies on how positive and negative mood affect executive functions in the brain from neuropsychology [45]. They are fundamentally important and can inform cognitive load measurement, but are not our focus in this study.

Recently, researchers have begun to use classification methods with eye data to measure cognitive load, but most work has been classifying two states or two levels and a few have classified three states, although classification among three or more affective states is often seen in emotion recognition research. Marshall used wavelet analysis for pupillary response together with blink number, saccade number and the difference between horizontal locations for the left and right eyes, and employed discriminant function and neural network analysis to investigate two cognitive states [8]. Those states, denoted as in-task state and in-rest states, focused and distracted states, alert and fatigue states, were classified in three different studies (30, 11 and 1 participant respectively). The best overall accuracies ranged from 69% to 92%. Another interesting work employed multiple modalities to classify three states of interest on participants while they were engaged in conversations. Eye movement and eye distance were used as a criterion of mental state and achieved an average accuracy of 45.6% alone [41]. Haapalainen et al. mapped six elementary cognitive tasks onto the three contextual factors: speed of closure, flexibility of closure and perceptual speed, and employed heat flux, ECG, GSR, median of pupil diameter, EEG and heart rate, with a Naïve Bayes classifier to discriminate between low and high levels [10]. Pupillary response performed better than GSR, close to EEG and heart rate but more poorly than ECG and heat flux. The average accuracy from pupil features achieved across 20 participants was 57.4% [10]. Clearly, the classification of multiple cognitive load levels, robustly in the presence of non-cognitive factors, is of interest.

Our work attempts to take the advantage of different eye activity patterns, select reliable features, process these with suitable measurement timing and differentiate more than two levels of cognitive load. Although we also used a controlled experiment, the emotion 'interference' provides a form of realism to the investigation of CLM.

3. Methodology

3.1. *Experiment setting*

In this study, cognitive load was induced using arithmetic tasks, and the difficulty level was controlled by the number of carries and digits. More details of the tasks can be found in Table 1. Emotional interference corresponding to different arousal and valence levels was induced by showing International Affective Picture System (IAPS) [31] images in the task background. The experiment was adapted from those using pupillary response for measuring cognitive load with arithmetic tasks [18,23] and for measuring arousal with IAPS images [21]. IAPS provides affective ratings from

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Table 1 – Descriptions of the five task difficulty levels and six emotion categories.

Task description: add four numbers	Four digits selected from	Difficulty level (95% confidence interval for the mean of the subjective rating (measured))	Performance score (measured)
No carry produced for each single number addition	{0,1}	1 (1.8 ± 0.4)	89.5%
A carry is produced in the lower digit in the third or fourth addition. One-digit and two-digit addition	{1,2,3,4,5}	2 (2.3 ± 0.4)	89.5%
A carry is produced in the lower digit for at least every second addition. One-digit and two-digit addition. The result has two digits	{5,6,7,8,9}	3 (2.9 ± 0.4)	98.1%
A carry is produced in the lower digit in the last one or two addition. Two-digit addition. The result has two digits	{10,11,12,13,14,15,16,17,18,19}	4 (4.0 ± 0.5)	77.6%
A carry is produced in both the low and high digit for every addition. Two-digit addition. The result has three digits	{84,85,86,87,88,89,90,91,92,93}	5 (6.1 ± 0.8)	56.2%

Image description	Emotion description	Arousal/valence (mean, SD of IAPS rating)
Aging, loneliness, dirty dishes, garbage	Boredom, sleepy	Low/negative (M = 3.63/3.73, SD = 2.02/1.37)
Crying victims, guns, vomit, skulls, roach	Disgust, sorrow	Medium/negative (M = 4.83/2.93, SD = 2.25/1.61)
Bloody faces, snake and spider, surgery, dying	Terror, distressed	High/negative (M = 6.02/2.88, SD = 2.16/1.72)
Animals, birds, landscape	Relaxed, peace	Low/positive (M = 3.51/6.91, SD = 2.21/1.45)
Animal families, smiling children, art performance	Joy, happiness	Medium/positive (M = 4.54/7.29, SD = 2.27/1.53)
Wedding, romance, adventurous sports, gold and money	Excitement, lust	High/positive (M = 5.90/7.16, SD = 2.17/1.53)

approximately 100 college students for a large set of photographs in three dimensions: valence, arousal and dominance. For the valence dimension, the rating from low to high indicates the range from negative to positive. For the arousal dimension, a low rating represents calm while a high rating means excited [31]. As valence and arousal are the two primary scales for emotional assessments, and for a comparison of studies in which only arousal and valence were used, we did not consider the dimension of dominance for emotion induction. Images from six categories were selected: low/high rating in valence and low/medium/high rating in arousal to elicit different emotions. The selected images and their average ratings from each category are shown in Table 1.

At the beginning of the task, the image was displayed, together with ten 'x' or 'xx' placeholders arranged in a circular pattern, for 2 s in order to allow the eye to adapt to the light intensity. Then four numbers were displayed sequentially and participants were required to sum the four numbers. Each number replaced one of the placeholders (selected randomly) and stayed for 3 s. When a number showed up, the previous number was replaced by 'x' or 'xx', so that the effect of light intensity change on pupillary response during the task was minimal. Displaying numbers in random placeholder positions encouraged the eye to explore image elements and engage with the content, to facilitate emotion induction. At the 15th second, 10 numbers, representing the candidate answers, replaced the 10 placeholders without the background changing. Participants were required to use a mouse to click the correct answer and then click a radio button in the center to submit the answer. The position of the correct answer for each addition task was changed randomly. After the addition task, a rating form for task difficulty (9-point scale from extreme easy to extreme difficult) or/and emotion (two dimensions in 3 scales) appeared and participants could only select one ratio button for each subjective assessment.

The participants comprised seven females and eight males, aged 20–48 (M = 26.8, SD = 7.2). After a short training, each participant firstly completed 1 task-centered session, with task demands and with gray color as background, then 1 image-viewing session without task demands but with 2

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Fig. 1 – Time line for each task. Each task comprises focusing, image viewing, reading and calculating four addends sequentially, selecting an answer and subjective rating of both task difficulty and emotion.

images from each emotion category. Then each participant completed 6 task-interference sessions, which contained 2 tasks in each of 5 cognitive load levels with 10 different images from the same emotion category. The sequence of the cognitive load levels was randomized to minimize carry-over effects. The sequence of emotion categories for the 6 sessions was low/positive, medium/positive, high/positive, and low/negative, medium/negative, high/negative in arousal/valence. At the end of each session, participants were advised to take a break as long as they needed to recover from images or/and cognitive tasks and then continue. Each task in the 1 task-centered and 6 task-interference sessions followed the same procedure as shown in Fig. 1. In the image-viewing session, each image was displayed for 14 s then an emotion rating form appeared. Therefore, a total of 82 recordings ($K = 82$) were obtained from each participant, including 60 samples with both cognitive load and emotion factors, 10 samples with only the cognitive load factor and 12 samples with only the emotion factor. The signal length of each sample was 14 s, during which four task stimuli were systematically presented and time stamped.

3.2. *Signal pre-processing*

Pupil dilation and position were monitored and recorded using a FaceLAB 4 [32] desk-mounted eye tracker system with a sampling rate of 60 Hz. Participants were free to move their head but instructed to keep their eyes within the screen display range. The 12 s of mental arithmetic activity containing the first to the last task stimuli is the time window of interest, and a pupil sample signal (P) was extracted according to the timestamps. Each recording of pupil size was first linearly interpolated during blinking and then low-pass filtered at around 4 Hz cut off frequency [33] to remove high frequency noise such as drift, tremors in eye and equipment introduced noise in the measure. In order to further attenuate the noise, we averaged left and right eye pupil diameters after filtering to obtain an averaged pupil signal of length $N = 720$.

Due to inaccurate blink detection from the eye tracker, measures of blink were processed from video recorded from a separate 30 Hz webcam situated in front of participants, around 0.5–0.7 m away. Scripts developed in MATLAB using motion analysis and template matching [34] were able to recognize eye blink states as either blink (1) or non-blink (0) during the 12 s time window of interest, which we denote as $B_k[n] \in \{1, 0\}$; $n = 1, \ldots, N/2$; $k = 1, \ldots, K$. Extracted blink data were superimposed on the video and played back to manually ensure that blink features extracted correctly represented actual blink activity.

Fixation and saccade data during the 12 s time window were extracted and obtained from pupil positions using a dispersion-based algorithm [25]. For each sample of pupil position, those eye positions remaining within 1 degree of visual angle for at least 100 ms were defined as fixations $F_k[n]$ with centroid position C (i.e. $F_k[n] \in \{C_j\}$, $n = 1, \ldots, N$; $k = 1, \ldots, K$, j is the index of the fixation). In the saccade vector, constant eye positions were recorded as fixations (0 s) and moving eye positions were recorded as saccades (1 s), i.e. $S_k[n] \in \{1, 0\}$; $n = 1, \ldots, N$; $k = 1, \ldots, K$.

3.3. *Measured feature sets*

We gathered a variety of features from the literature and also proposed some features to investigate: zero crossing count of pupil size, features from cumulative blink/fixation/saccade number, eye features from task stimuli onset to the first saccade. These were calculated in different segments 1 s (spanning from 0.5 s before to 0.5 s after addend), 1.5 s (spanning 1.5 s after addend), 2 s (spanning 0.5 s after addend to 0.5 before next addend), 3 s (spanning 3 s after addend) around the four task stimuli ('1S', '2S', '3S', '4S', where 'S' is the stimulus). The segment names are added to each feature name as suffixes to denote the segmentation scheme. This was to evaluate which task stimulus produces a significant response from eye features in each difficulty level with emotion interference; and how soon and for how long eye features should be measured after each task stimulus. Long delays after task stimuli might weaken the cognitive load effect, while short segments might result in low signal-to-noise ratios.

Throughout the paper, $k \in \{1, \ldots, K\}$ is the number of tasks that each participant has completed. Each feature was calculated on a per-task basis. For simplicity, k is omitted in the equations below. $n \in \{0, \ldots, N-1\}$ is the discrete-time sample index, and n_1 and n_2 ($n_1 < n_2$) are the first and last sampling

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

indices of a particular segment. The features extracted are defined as follows.

The *pupil diameter change*:

$$PD[n] = P[n] - PB[n] \tag{1}$$

where *PB* is the average baseline during *1S1sec*;

The *cumulative blink number*:

$$BD[n] = \sum_{n=n_1}^{n_2} B[n]; \tag{2}$$

The *blink number* during $[n_1, n_2]$:

$$BNum = \frac{1}{2}\sum_{n=n_1}^{n_2} |B[n] - B[n-1]|; \tag{3}$$

The *blink duration per blink* during $[n_1, n_2]$:

$$BDur = \frac{1}{BNum}\sum_{n=n_1}^{n=n_2} B[n]; \tag{4}$$

The *cumulative saccade number*:

$$SD[n] = \sum_{n=n_1}^{n_2} S[n]; \tag{5}$$

The *fixation number* during $[n_1, n_2]$:

$$FNum = \frac{1}{2}\sum_{n=n_1}^{n_2} |\bar{S}[n] - \bar{S}[n-1]|, \bar{S}[n] = 1 - \bar{S}[n]; \tag{6}$$

The *fixation duration per fixation* during $[n_1, n_2]$:

$$FDur = \frac{1}{FNum}\sum_{n=n_1}^{n_2} \bar{S}[n]; \tag{7}$$

The *saccade amplitude per saccade* during $[n_1, n_2]$: C_x and C_y are the centroid positions of fixation $F[n]$ along the x and y axes.

$$SAmp = \frac{1}{FNum - 1} \times \sum_{j=2}^{J} (\max(C_x)) - \min(C_x) + \max(C_y) - \min(C_y)); \tag{8}$$

Based on the above features (raw features), we then calculated their mean, standard deviation and difference values during $[n_1, n_2]$. We add a '−*M*' suffix to the raw feature name (*fn*), for example, *PDM*[*n*], *BDM*[*n*], *BDurM*[*n*], *SDM*[*n*], *FDurM*[*n*], *SAmpM*[*n*], to denote the means of the raw features. Similarly we add the *−Std*, *−Diff1*, *−Diff2* suffices to denote the standard deviation, average of absolute difference value and average of slope of the raw features.

$$fnM = \frac{1}{n_2 - n_1}\sum_{n=n_1}^{n_2} \text{raw feature}[n]; \tag{9}$$

$$fnStd = \sqrt{\frac{1}{n_2 - n_1 - 1}\sum_{n=n_1}^{n_2} (\text{raw feature}[n] - fnM)^2}; \tag{10}$$

$$fnDiff1 = \frac{1}{n_2 - n_1}\sum_{n=n_1}^{n_2} |\text{raw feature}[n] - \text{raw feature}[n-1]|; \tag{11}$$

$$fnDiff2 = \frac{1}{n_2 - n_1}\sum_{n=n_1}^{n_2} (\text{raw feature}[n] - \text{raw feature}[n-1]); \tag{12}$$

In addition, we calculated the zero crossing count of *PD*, after filtering by a low pass filter.

$$ZCC = \frac{1}{2}\sum_{n=n_1}^{n_2} |sign(PS[n]) - sign(PS[n-1])|, \quad PS[n] = PD[n] - PDM; \tag{13}$$

In addition to the above features, we explored how *PD*, *BDur*, *FDur* and *SAmp* varied in response to the task stimulus, that is, during the time from *1S*, *2S*, *3S*, *4S* onset to the next saccade occurrence. Therefore, another four features, the means of *PD*, *BDur*, *FDur* and *SAmp* during that time, denoted by *mSS* (*m*th task stimulus to the next saccade), were calculated: *PDMmSS*, *BDurmSS*, *FDurmSS*, *SAmpmSS*.

3.4. Statistical analysis

When measuring pupil response change, it was not easy to select a baseline segment during which light intensity was always identical to that during tasks. We used the average pupil size during the 0.5 s before and after the first task stimulus (i.e. 1.5th to 2.5th second) as the baseline and assumed that background luminance was constant afterwards. Some rapid changes as a result of light contrasts during saccades were filtered in the signal pre-processing stage. During the first task stimulus, there was no addition needed and participants should have experienced minimal cognitive load. In the image-viewing session, since each participant possibly perceived different emotional stimuli within a single image and the precise time(s) of this response was unknown, we set the average pupil size during the first half second after image onset as the baseline.

A three-way repeated ANOVA test, followed by Bonferroni corrected *t*-tests, was conducted for each eye feature in the task-centered, image-viewing and task-interference sessions. The aim was to find the sensitive features and their measurement timing across all participants and evaluate the cognitive load effect, emotion effect and their interactions for each feature. We also considered ω^2 values, which are an

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

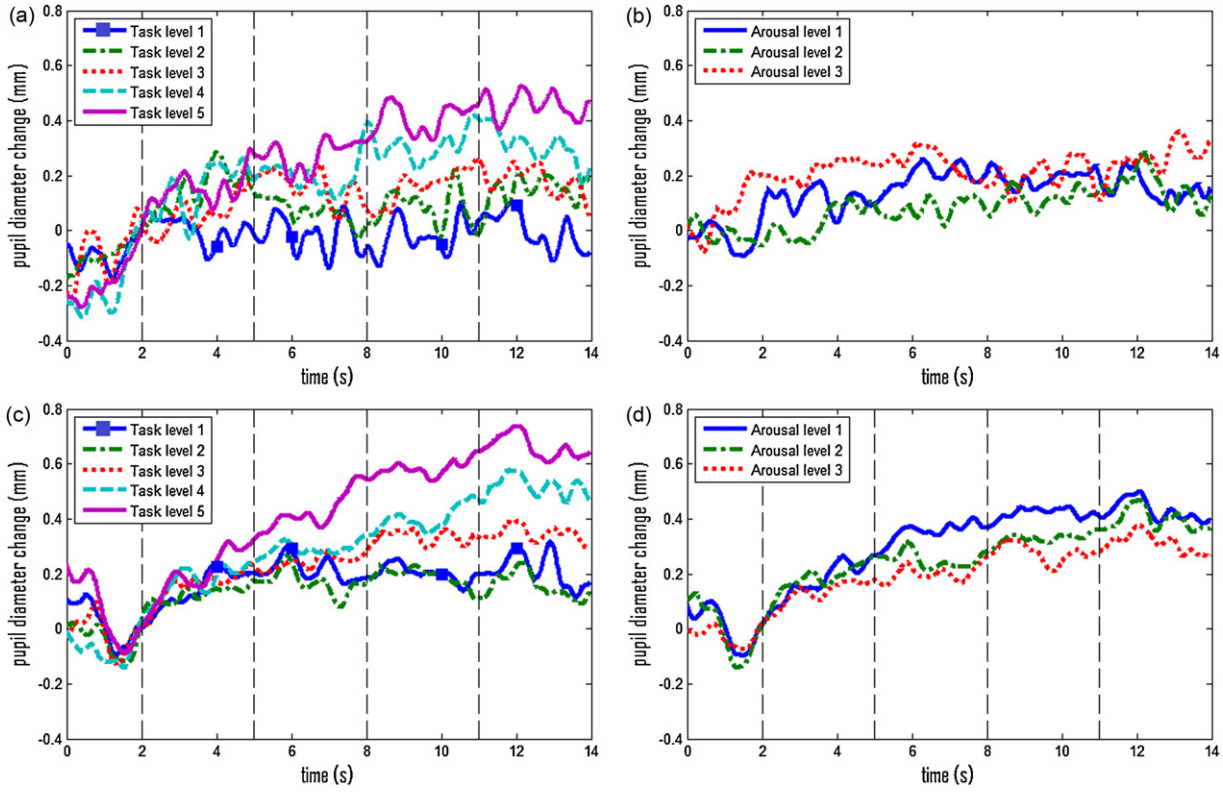


Fig. 2 – Comparisons of pupillary response *PD*[*n*] during (a) the task-centered sessions, (b) the image-viewing sessions and (c), (d) the task-interference sessions, averaged across all 15 participants. (c) and (d) are from the same data but averaged across all arousal levels and task levels respectively.

unbiased measure of the effect size in ANOVA. These indicate the magnitude of variation in eye features that is explained by cognitive load or arousal factor. The three-factor design was task (5 levels) × arousal (3 levels) × valence (2 levels). We set 0.05 as the critical p value and for those within-subject tests that violated the assumption of sphericity, the degrees of freedom were corrected by Greenhouse–Geisser epsilon coefficients.

In ANOVA tests, the p value can only determine whether the observed value of a statistic differs sufficiently from a hypothesized value of a parameter to draw the inference. To understand the exact weights of the significant contributions from cognitive load and from arousal factors in pupil features, we conducted multiple regression analysis to evaluate the measurement timing. We applied the ANOVA model on each eye feature to look into the cognitive load effect in the task-centered session, the arousal effect in the image-viewing session, and the cognitive load, arousal and cognitive load × arousal in the task-interference sessions respectively. For the regression analysis, we used all the three sessions at a time to estimate the extent to which cognitive and arousal factors predict pupillary response and to find the segment across all participants with the highest weight for the cognitive load effect and the lowest weight for the arousal effect. The model we used was:

$$Y = \beta_t X_t + \beta_a X_i + C, \quad (14)$$

where Y is the *PD* in the task-interference session (15×1 vector of 3 arousal levels and 5 task difficulty levels for each participant), X_t is the *PD* from the task-centered session (5×1 vector comprising 5 task difficulty levels, replicated for each emotion category for each participant) and X_i is the *PD* from the image-viewing session (3×1 vector comprising 3 arousal levels, replicated for each task difficulty level for each participant).

In applying this analysis, *PD* data from the two task repetitions were averaged, and similarly *PD* data from images from the same arousal-valence category from each participant were averaged. The coefficients were standardized in order to compare them on the same scale.

4. Statistical results

4.1. *Pre-processing for pattern analysis*

As a preliminary investigation, we examined the eye activity responses to the task-centered, image-viewing and task-interference sessions respectively, averaged over all participants. Figs. 2 and 3 present the measurable pattern changes of pupillary response and cumulative blink number respectively along the task timeline for the task-centered (Figs. 2 and 3(a)), image-viewing (Figs. 2 and 3(b)) and task-interference sessions (Figs. 2 and 3(c), (d)). The descriptive results are shown in Table 2.

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

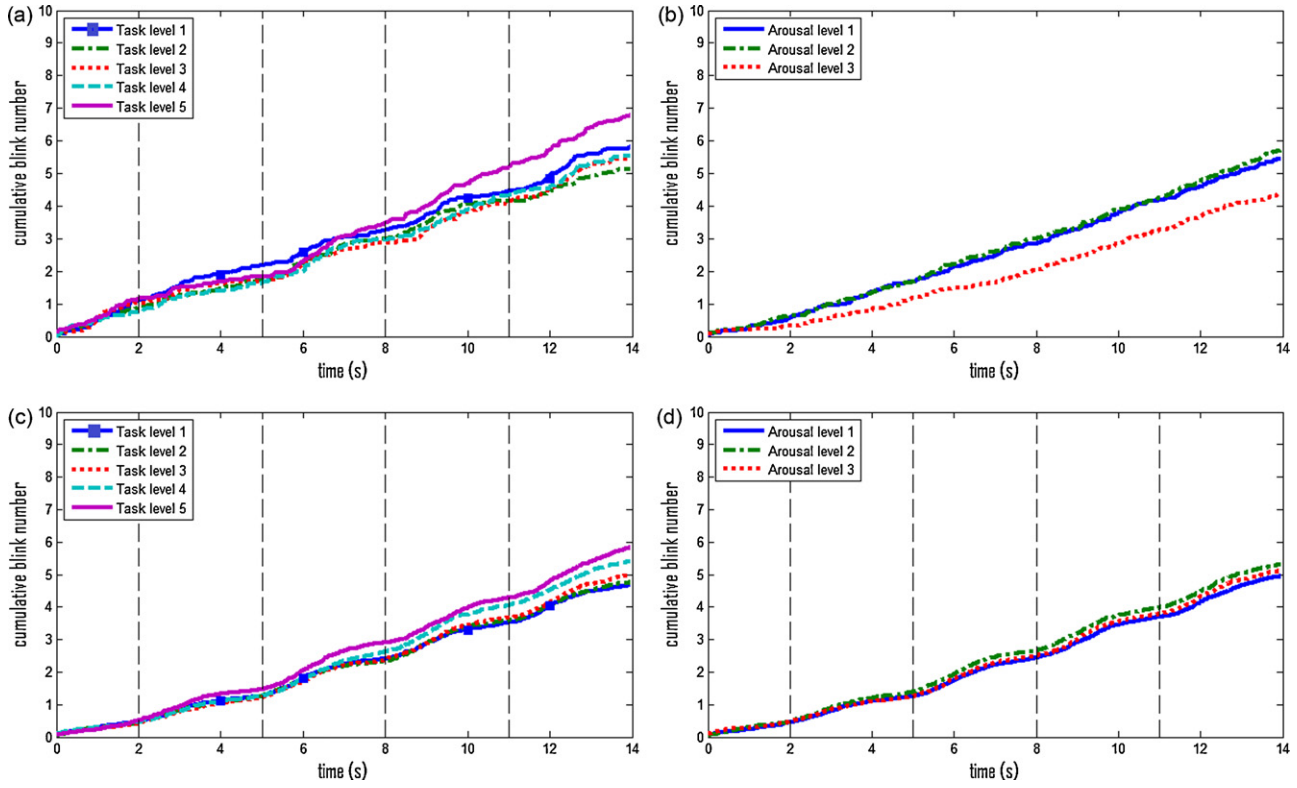


Fig. 3 – Comparisons of cumulative blink number *BD[n]* during (a) the task-centered sessions, (b) the image-viewing sessions and (c), (d) the task-interference sessions, averaged across all 15 participants. (c) and (d) are from the same data but averaged across all arousal levels and task levels respectively.

Table 2 – Descriptive statistical results for the three eye features in the three sessions. Mean and standard deviation values are presented in the form of 'mean (standard deviation)', with the order from low to high for task difficulty and arousal levels, negative and positive for valence levels.

	Task-centered session	Image-viewing session	Task-interference sessions
Pupil size (*PDMS12sec*) (difference to baseline in mm)	5 task difficulty levels: −0.01(0.16), 0.11(0.18), 0.14(0.21), 0.24(0.26), 0.32(0.26)	3 arousal levels: 0.18(0.18), 0.06(0.23), 0.26(0.28) 2 valence levels: 0.16(0.24), 0.18(0.21)	5 task difficulty levels: 0.19(0.11), 0.16(0.12), 0.27(0.15), 0.34(0.13), 0.46(0.18) 3 arousal levels: 0.34(0.12), 0.29(0.14), 0.23(0.11) 2 valence levels: 0.24(0.13), .33(0.12)
Blink number (*BNumS12sec*) (number)	5 task difficulty levels: 4.77(3.21), 4.30(2.48), 4.47(3.40), 4.80(3.46), 5.70(3.94)	3 arousal levels: 4.93(4.06), 5.13(3.36), 4.07(2.90) 2 valence levels: 4.78(3.65), 4.64(3.23)	5 task difficulty levels: 4.24(2.68), 4.34(2.72), 4.57(2.82), 4.97(3.47), 5.37(3.63) 3 arousal levels: 4.55(2.86), 4.86(3.00), 4.69(3.18) 2 valence levels: 4.74(2.99), 4.65(3.02)
Saccade amplitude (*SAmpS12sec*) (cm/saccade occurrence)	5 task difficulty levels: 1.16(0.95), 1.31(0.96), 1.24(1.02), 1.60(1.27), 1.99(1.21)	3 arousal levels: 1.03(0.71), 1.36(0.97), 1.02(0.78) 2 valence levels: 1.11(0.76), 1.16(0.82)	5 task difficulty levels: 1.09(0.85), 1.00(0.73), 1.05(0.70), 1.18(0.84), 1.24(0.90) 3 arousal levels: 1.17(0.84), 1.14(0.73), 1.03(0.88) 2 valence levels: 1.00(0.81), 1.22(0.81)

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Table 3 – ANOVA results for the three eye features in the three sessions. Levels in brackets with '–' between are significantly different paired levels using Bonferroni corrected t tests. The threshold is 0.05/number of comparison pairs (10 for five levels and 3 for three levels).

	Task-centered session	Image-viewing session	Task-interference sessions
Pupil size (*PDMS12sec*)	$F(3.0,41.2) = 4.58$, $p = 0.01$, (1–4, 1–5);	Arousal: $F(2,28) = 4.64$, $p = 0.02$, (2–3); Valence: $F(1,14) = 0.18$, $p = 0.67$; Arousal × Valence: $F(2,28) = 2.41$, $p = 0.11$;	Task: $F(2.9,40.2) = 19.99$, $p < 0.01$, (1–4, 1–5, 2–4, 2–5, 3–5, 4–5); Arousal: $F(1.8,25.1) = 8.41$, $p < 0.01$, (1–3); Valence:: $F(1,14) = 6.20$, $p = 0.03$; Task × Arousal: $F(5.1,72.0) = 3.43$, $p < 0.01$ Task × Valence: $F(2.9,40.2) = 2.14$, $p = 0.11$ Arousal × Valence: $F(1.8,25.1) = 6.62$, $p < 0.01$
Blink number (*BNumS12sec*)	$F(2.2,30.5) = 2.4$, $p = 0.10$; (2–5, $p = 0.006$)	Arousal: $F(1.4,19.4) = 4.17$, $p = 0.04$, (2–3); Valence: $F(1,14) = 0.11$, $p = 0.74$; Arousal × Valence: $F(1.4,19.4) = 6.91$, $p = 0.01$;	Task: $F(1.5,21.4) = 4.39$, $p = 0.03$, (1–5, 2–5); Arousal: $F(1.9,26.9) = 1.88$, $p = 0.17$; Valence:: $F(1,14) = 0.38$, $p = 0.55$; Task × Arousal: $F(2.9,41.1) = 1.31$, $p = 0.28$ Task × Valence: $F(1.5,21.4) = 0.79$, $p = 0.43$ Arousal × Valence: $F(1.9,26.9) = 6.91$, $p < 0.01$
Saccade amplitude (*SAmpS12sec*)	$F(3.0,41.7) = 4.29$, $p = 0.01$, (1–5, 2–5, 3–5);	Arousal: $F(2,28) = 3.03$, $p = 0.06$; Valence: $F(1,14) = 0.14$, $p = 0.72$; Arousal × Valence: $F(2,28) = 0.55$, $p = 0.58$;	Task: $F(2.7,38.0) = 3.09$, $p = 0.04$, (2–5); Arousal: $F(1.8,26.0) = 1.07$, $p = 0.35$; Valence: $F(1,14) = 4.84$, $p = 0.05$; Task × Arousal: $F(5.0,70.4) = 0.52$, $p = 0.76$ Task × Valence: $F(2.7,38.0) = 1.80$, $p = 0.17$ Arousal × Valence: $F(1.8,26.0) = 0.55$, $p = 0.58$

4.2. *ANOVA model*

Subjective rating of task difficulty/emotion was used as a reference to indicate the degree of distinction between the levels for all participants. The average scale values of the difficulty levels and their confidence intervals can be found in Table 1. Repeated ANOVA results ($F(1.6,21.9) = 91.4$, $p < 0.01$) showed that tasks were felt to be more difficult by the participants when the difficulty levels increased. The pairs of levels 1–3, 1–4, 1–5, 2–4, 2–5, 3–4, 3–5, 4–5 were found significantly different using a Bonferroni corrected *t*-test. Meanwhile, subjective ratings for the three arousal levels ($M = 1.55, 1.88, 2.24$; $SD = 0.24, 0.18, 0.26$) showed participants perceived increasing arousal when they were viewing more arousing images selected from IAPS ($F(1.9,26.9) = 91.9$, $p < 0.01$) and the three induced arousal levels all yielded significant differences by Bonferroni corrected *t*-test. The differences in subjective rating between the two levels of valence ($M = 1.37, 2.59$; $SD = 0.17, 0.19$) were also significant ($t(14) = 17.6$, $p < 0.01$), in accordance with the levels of the selected IAPS images.

Tables 2 and 3 show only the most effective eye features during the 12 s time window to build a general view of which features were relevant to cognitive load, arousal factor and their interactions. When examining the measurement timing, we found that the features exhibiting all three significant effects, arousal, cognitive load and arousal × cognitive load, were all pupil diameter averages for segments around the second to the last task stimulus (*PM2S1sec*, *PM3S1sec*, *PM4S1sec*; *PM2S1.5sec*, *PM3S1.5sec*, *PM4S1.5sec*; *PM2S2sec*, *PM3S2sec*, *PM4S2sec*; *PM2S3sec*, *PM3S3sec*, *PM4S3sec*; *PM2SS*, *PM3SS*, *PM4SS*). An interesting trend is that as more task stimuli were presented in this experiment (requiring sustained effort), pupil size increased, meanwhile, the ω^2 value on cognitive load effect increased from 0.06 after the 1st stimulus to 0.23 after the 4th stimulus but the ω^2 value on arousal effect decreased from 0.03 to 0.01. Therefore, around the last task stimulus, which was the most difficult part of the task (this part most accurately represents the designed difficulty levels), the average pupil size reached its peak with the maximum effect from cognitive load and minimum effect from arousal effect. This trend agrees with a previous study [23] in that arousal effect was obvious when cognitive load is minimal. However, this trend was not observed on the blink and eye movement features.

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Table 4 – Selected features ranked by a ratio of weights in multiple regression analysis.

Features	β_t	β_a	$\lvert\beta_t/\beta_a\rvert$
PDM4S1sec	**0.3557**	**0.0128**	**27.7**
PDM4S1.5sec	**0.34387**	**−0.0139**	**24.8**
PDM4S3sec	**0.3558**	**0.01673**	**21.3**
PDM4S2sec	0.3533	0.01930	18.3
PDM2H6sec	0.3292	0.01808	18.2
PDM4S2sec	0.1766	0.01835	9.6
PDM3SS	0.1545	0.02140	7.2

The best three ratios are shown in bold.

Among the three types of eye activity, pupil features demonstrated the best attributes for discriminating different levels of cognitive load. *PM4S1.5sec* and *PM4S3sec* obtained the largest ω^2, followed by the derived features *ZCC* and *PDdiff1*. Blink was the second most promising type of eye activity for CLM, in terms of its number of features having significant effects on cognitive load, although two of the participants had very low blink numbers (0–2) in the five cognitive load levels. Among the eye movement features, only *SAmpS12sec* and *FDur4S1sec* showed potential for discriminating different load levels.

Another major concern is to what extent the pupil light reflex was reduced by the baseline subtraction when participants were free to explore the non-uniform image background with task stimuli. A two-way repeated ANOVA test for *PDM4S3sec* on the luminance[1] (0–1) of images (4 levels) and cognitive load (5 levels) showed that before baseline subtraction, pupil size had significant effects on the luminance of images, cognitive load levels and the interaction of the luminance and cognitive load; however, after baseline subtraction, only the cognitive load effect survived, which suggested that the baseline was reasonable. Still pupil size might be affected by the local luminance of image due to gaze shifts, which will cause large variations within the 5 load levels. In a comparison of the standard deviations of *PDMS12sec* after baseline subtraction within the 5 load levels in the task-centered session (gray background) and task-interference sessions (image background), we found that they are not significantly different. This suggested that pupil light reflex due to gaze shifts did not dominate over cognitive load effect but might undermine cognitive load estimation as noise.

4.3. *Multiple regression analysis for pupil features*

From ANOVA results, only pupil features in different segments showed significant effects on both cognitive load and arousal factors. Examining the results from the multiple regression analysis for pupillary response in Table 4, we found the ratios for *4S1sec*, *4S1.5sec* and *4S3sec* were very close and had a higher contribution from cognitive load. Considering that pupil diameter in the *4S3sec* segment had the largest ω^2 in the task-inference session, we chose the segment of *4S3sec* for the average pupil size in CLM.

4.4. *Deciding the candidate feature set for classification*

To find a feature set for classification, we chose candidate features from all proposed eye features based on their p values, ω^2 values, origins in different types of eye activity, and consistency of differentiating load levels in task-centered and task-interference sessions. The features that best met the above criteria were *PDM4S3sec*, *BNumS12sec*, *ZCC4HzS12sec*, *PDdiff14S3sec*, *BDdiff1S12sec*, *SAmpS12sec* and *FDur4S1sec*.

5. Classification results

5.1. *Gaussian mixture model classifier*

Gaussian mixture models (GMMs) were selected for classification purposes because of their suitability for modeling arbitrary feature distributions and the explicit control over the number of model parameters given the small size of the database. In GMM classification, the probability density function (pdf) parameters, comprising the means (m), covariances (C) and weights (w) are estimated based on a training dataset.

$$p(X) = -\sum_{m=1}^{M} W_m \frac{1}{(2\pi)^{k/2}|C_m|^{1/2}} \exp\left(-\frac{1}{2}(X-\boldsymbol{m}_m)^T C_m^{-1}(X-\boldsymbol{m}_m)\right); \quad (15)$$

where m is the number of mixture components, and k is the dimension of feature vector. The iterative Expectation Maximization (EM) algorithm is usually used to maximize the likelihood of data distribution for the mixture components. Because of the limited samples for training, we fixed the number of mixture components to1.

Individual models were built for each participant (subject-dependent classification) and one GMM was trained for each of the five load levels. During testing, the GMM with the highest log likelihood among test vectors was chosen and the test vector was assigned the label of the load level for that Gaussian model. Classification was conducted using a leave-one-session-out average over 7 folds, where 10 tasks from one of the 7 sessions were used as test data and the remaining data were used to train the classifier. Training and testing features were normalized to have zero mean and a standard deviation of one on a per-feature basis before classification.

6. Results

The above candidate features give good performance for most individual participants. Inevitably there are some participant-dependent features. Among the seven selected features, *PDM4S3sec* and *BNumS12sec* demonstrated consistent trends for cognitive load in both task-centered and task-interference sessions and they were also the most frequently reported features that were correlated with cognitive load in different tasks; hence we used *PDM4S3sec* and *BNumS12sec* as core features in all classification experiments, and used a filter method to select participant-dependent features among the remaining

[1] To calculate the luminance of an image, we firstly converteda RGB image to an intensity image by 0.2989 × Red + 0.5870 × Green + 0.1140 × Blue to obtain the value of each pixel,and then averaged them.

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Table 5 – Cognitive load classification accuracy using pupil dilation (PD), blink (B) and other eye activity features with a GMM approach. Levels in brackets with ',' between are grouped into one class and levels with '–' between are distinct classes.

Classes (%)	*PDM4S3sec* + *BNumS12sec*	*PDM4S3sec* + *BNumS12sec* + *participant-dependent features*	Best individual accuracy	Worst individual accuracy	Reaction time (RT)	Best accuracy for RT	Worst accuracy for RT
2 classes (1,2–4,5)	69.3	71.1	94.6	48.2	74.4	89.3	60.7
3 classes (1–4–5)	49.4	50.3	66.7	28.6	55.9	73.8	35.7

five features, to try to better understand the individual upper performance bound.

The Fisher ratio J_i was calculated for each feature in each class i using a one-against-all method without projection, and used to select the features with maximum value in the 1st to ith class. This only applied to features excluding *PDM4S3sec* and *BNumS12sec*, therefore the number of features used for classification ranged from 2 (*PDM4S3sec* and *BNumS12sec*) to $2+i$ (selected i times) depending on the training data and number of classes.

GMM classification accuracy was scored by comparing the predicted load levels with the induced cognitive load levels. Since the statistical tests for subjective ratings of task difficulty showed that three levels are significantly different from each other, we used the verified three levels as the ground truth. Table 5 presents the average classification accuracies across 15 participants. The classification accuracy varies from participant to participant; therefore we also list the best and the worst classification performance. We demonstrate 2- and 3-class results by selecting the distinguishable levels, yielding some insight into the trade-off between the number of cognitive load levels (precision) and accuracy. A comparison with the reaction time measure is also presented for reference.

7. Discussion and conclusion

7.1. *Results for cognitive load measurement*

ANOVA test results in Section 4.2 and multiple regression results in Section 4.3 revealed important implications for using eye activity for CLM. Pupil size [7,12,13,18] and blink number [11,19] increased with more difficult tasks, which perfectly matches the literature. Pupil size also increased with higher arousal images regardless of valence, which is also consistent with studies of pupil dilation using visual [21] and auditory stimuli [22]. However, pupil size increased with images of positive valence when a task goal was presented in this study, as the p value was close to 0.05.

The new finding here is that some eye activity feature patterns (notably pupil dilation and blink) for the cognitive load levels were not significantly altered with or without arousal factor in the task-goal driven situation. In contrast, the patterns of features for the arousal level seemed weakened in the pupillary response when cognitive load was induced and there was no arousal effect on the features of blink, fixation and saccade. This result suggests the dominance of cognitive load over emotion in eye features during task performance. Although a previous study only observed an arousal effect in pupillary response when cognitive load was very low with auditory stimuli [23], our work still found an arousal effect in high cognitive load tasks. Since pupillary response showed a dependence on arousal factor, different possible segmentation approaches were considered to reduce this effect. *PDM4S3sec* seemed a more suitable feature than *PDMS12sec* for CLM, exhibiting a high ratio of cognitive load effect to arousal factor. Meanwhile there is no sufficient evidence supporting the dependence of cognitive load on arousal factor for the features

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

of blink, fixation and saccade, as there were no significant effects in the cognitive load × arousal factor interaction.

With the two core features identified, *PDM4S3sec* and *BNumS12sec*, together with other eye features selected on individual bases by Fisher criterion we achieved an overall classification accuracy of 71.1% for two levels and 50.3% for three levels across all 15 participants under the presence of interference from emotion. The results are close to the accuracy produced by the reaction time measure, which is a major measurement for clinical cognitive test. Comparing eye-feature based load level estimation with reaction time, the advantages of the former are that it does not require any conscious user (patient) response to a stimulus, does not require actions to respond, and importantly can be measured continuously through a task.

An average of 70% classification accuracy for two classes of cognitive load is also better than another initial effort to classify two-level cognitive load using pupillometry, with an average accuracy of 57% [10]. Our classification accuracy was achieved in the presence of emotional interference, while most results from the literature were obtained by less realistic 'pure' cognitive load tasks. It is difficult to directly compare our results with Marshall's work, where the overall two-class accuracy ranges from 69% to 92% [8] for classifying two different cognitive states (e.g. working and resting) instead of between cognitive load levels, since our low, medium and high cognitive load levels are ordinal and implying a ranking between them. The encouraging results are that two load levels can be distinguished effectively, and that pupillary and blink features are robust in the presence of another form of arousal. The most distinguishable load levels in this study are levels 1, 4 and 5. Accuracies for other groups can be reduced by 3% at most. Therefore, our work sets a benchmark for CLM under non-ideal task conditions, achieving similar accuracy to the reaction time measures and showing the possibility for CLM in near-real time.

It is worth noting that the primary objective of this work was to investigate the interaction between cognitive load and arousal in eye activity features rather than to optimize the system's classification accuracy. This is because reducing ambiguity as to what an eye activity based system is measuring can be considered an essential precursor to developing practical CLM systems. In this experiment, task difficulty levels were randomized and the sequence was the same for every participant. However counterbalancing the difficulty levels may be a better design. In the repeated ANOVA tests on the differences of pupil size, blink number and saccade amplitude in the task-centered, image-viewing and task-interference sessions, we found that there was no evidence to support an ordering effect due to the ordering of difficulty level presentation, such as fatigue, on pupil size, blink number and saccade amplitude. This suggests that the randomization of levels did not distort the results presented above.

7.2. *Challenges for cognitive load measurement*

Clearly the precision of eye-based CLM remains a general challenge. However, the best accuracies revealed in Table 5 show some promise. Understanding the disparity between the best and worst performing participants may provide a viable path toward improvements. Firstly, the accuracy of recording eye activity using an eye tracker depends heavily on the calibration procedure. Low quality calibration brings inaccurate models for participants. Too much head movement and frequent changes in the distance from the head to the screen also resulted in inaccurate eye tracking data. Noise, such as light reflex, can be minimized by averaging during ANOVA and regression analysis, but it significantly affected the classification accuracy when each sample was tested. Habituation effects may be another limitation in this study, as the number of repetitions of the three sessions was not equal. Future work will look into the specific eye activity measurements, with a view to measuring pupil dilation in particular, directly from the raw video image rather than depending on a third-party tracker that has not been optimized for CLM purpose. Another tentative reason for low accuracy could be the varying cognitive capacities among participants [28]; there may be different distances between levels across different individuals.

7.3. *Implications for realistic applications of cognitive load measurement*

By controlling certain aspects of the experiments herein, some limitations for realistic applications are implied. In practice, we would need to know the time of the first task stimulus to obtain a baseline for pupillary response, which is subsequently subtracted to normalize for the light effect on the eye. We also need a stimulus for measurement, which we term the 'test marker', in our approach. The test marker can be either a task stimulus or can be task related, e.g. the arrival of a target to hit. The time from the first task stimulus to the 'test marker' can be seen as the temporal resolution for capturing dynamic cognitive-load-related changes and cannot be too short if blink features are to be used. Light is another noise source for pupillary response and needs to be controlled at least during the measurement periods. Blink, saccade and fixation can be monitored from the first task stimulus to the 'test marker'. In the context of realistic applications, we envisage an outwards facing camera or light sensor to measure ambient light. If we know the pupil size at a certain luminance, we can reduce the effect of light reflex by subtracting the pupil size that corresponds to the luminance where the gaze is directed [43].

Acknowledgments

The authors would like to thank S. Hussain from the University of Sydney for assisting with experiment design and data collection. We also thank anonymous reviewers for their valuable comments.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

REFERENCES

[1] A. Collie, D. Darby, P. Maruff, Computerised cognitive assessment of athletes with sports related head injury, British Journal of Sports Medicine 35 (2001) 297–302.
[2] W.D. Spaulding, S.K. Fleming, D. Reed, M. Sullivan, D. Storzbach, M. Lam, Cognitive functioning in schizophrenia: implications for psychiatric rehabilitation, Schizophrenia Bulletin 25 (1999) 275–289.
[3] J. Brandt, N. Butters, C. Ryan, R. Bayog, Cognitive loss and recovery in long-term alcohol abusers', Archives of General Psychiatry 40 (1983) 435–442.
[4] R.C. Petersen, G.E. Smith, S.C. Waring, R.J. Ivnik, E.G. Tangalos, E. Kokmen, Mild cognitive impairment: clinical characterization and outcome, Archives of Neurology 56 (1999) 303–308.
[5] L.P. Sands, K. Yaffe, K. Covinsky, M.-M. Chren, S. Counsell, R. Palmer, R. Fortinsky, C.S. Landefeld, Cognitive screening predicts magnitude of functional recovery from admission to 3 months after discharge in hospitalized elders, The Journals of Gerontology: Series A 58 (2003) 37–45.
[6] G.F. Wilson, F.T. Eggemeier, Psychophysiological assessment of workload in multi-task environments, in: D.L. Damos (Ed.), Multiple-task Performance, Taylor & Francis Ltd., London, 1991, pp. 329–360.
[7] S. Chen, J. Epps, N. Ruiz, F. Chen, Eye activity as a measure of human mental effort in HCI, in: Proceedings of the 2011 Intelligent User Interface, Palo Alto, USA, 2011, pp. 315–318.
[8] S.P. Marshall, Identifying cognitive state from eye metrics, Aviation, Space and Environmental Medicine 78 (2007) 165–175.
[9] J. Klingner, B. Tversky, P. Hanrahan, Effects of visual and verbal presentation on cognitive load in vigilance, memory and arithmetic tasks, Psychophysiology 48 (2011) 323–332.
[10] E. Haapalainen, S. Kim, J.F. Forlizzi, A.K. Dey, Psycho-physiological measures for assessing cognitive load, in: Proceedings of the UbiComp' 2010, 2010, pp. 301–310.
[11] J.A. Veltman, A.W.K. Gaillard, Physiological workload reactions to increasing levels of task difficulty, Ergonomics 41 (1998) 656–669.
[12] K.F. Van Orden, W. Limbert, S. Makeig, T.P. Jung, Activity correlates of workload during a visuospatial memory task, Human Factors: The Journal of the Human Factors and Ergonomics Society 43 (2001) 111–121.
[13] T. de Greef, H. Lafeber, H. Oostendorp, J. Lindenberg, Eye movement as indicators of mental workload to trigger adaptive automation, in: D. Schmorrow, I. Estabrooke, M. Grootjen (Eds.), Foundations of Augmented Cognition, Neuroergonomics and Operational Neuroscience, Springer, 2009, pp. 219–228.
[14] R.W. Picard, E. Vyzas, J. Healey, Toward machine intelligence: analysis of affective physiological state, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 1175–1191.
[15] J. Kim, E. Andre, Emotion recognition based on physiological changes in music listening, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 2067–2083.
[16] R.J. Larsen, B.L. Fredrickson, Measurement issues in emotion research, in: D. Kahneman, E. Diener, N. Schwarz (Eds.), Well-being: The Foundations of Hedonic psychology, Russell Sage Foundation, NY, 1999, pp. 40–60.
[17] D.E. Irwin, L.E. Thomas, Eyeblinks and cognition, in: V. Coltheart (Ed.), Tutorials in Visual Cognition, Psychology Press, Talyor & Francis Group, New York, London, 2010, pp. 121–141.
[18] J. Beatty, Task-evoked pupillary responses, processing load, and the structure of processing resources, Psychological Bulletin 91 (1982) 276–292.
[19] Y. Tanaka, K. Yamaoka, Blink activity and task difficulty, Perceptual and Moto Skills (1993) 55–66.
[20] A.F. Kramer, Physiological metrics of mental workload: a review of recent progress, in: D.L. Damos (Ed.), Multiple-task Performance, Taylor & Francis Ltd, London, 1991, pp. 279–328.
[21] M.M. Bradley, L. Miccoli, M.A. Escrig, P.J. Lang, The pupil as a measure of emotional arousal and autonomic activation, Psychophysiology 45 (2008) 602–607.
[22] T. Partala, V. Surakka, Pupil size variation as an indication of affective processing, International Journal of Human-Computer Studies 59 (2003) 185–198.
[23] R.F. Stanners, M. Coulter, A.W. Sweet, P. Murphy, The pupillary response as an indicator of arousal and cognition, Motivation and Emotion 3 (1979) 319–340.
[24] E. Ponder, E.P. Kennedy, On the act of blinking, Experimental Physiology 18 (1927) 89–110.
[25] D.D. Salvucci, J.H. Goldberg, Identifying fixations and saccades in eye-tracking protocols, in: Proceedings of the 2000 Symposium on Eye Tracking Research and Applications, New York, 2000, pp. 71–78.
[26] A.T. Duchowski, Eye Tracking Methodology: Theory and Practice, 1st ed., Springer, 2003.
[27] R.J.K. Jacob, K.S. Karn, Eye tracking in human-computer interaction and usability research: ready to deliver the promises, in: Hyona, Radach, Deubel (Eds.), The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, Oxford, England, 2003.
[28] E.A. Byrne, R. Parasuraman, Psychophysiology and adaptive automation, Biological Psychology 42 (1996) 249–268.
[29] S. Ahern, J. Beatty, Pupillary responses during information processing vary with scholastic aptitude test scores, Science 205 (1979) 1289–1292.
[30] T. Piquado, D. Isaacowitz, A. Wingfield, Pupillometry as a measure of cognitive effort in younger and older adults, Psychophysiology 47 (2010) 560–569.
[31] P.J. Lang, M.M. Bradley, B.N. Cuthbert, International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual, Technical Report A-8, University of Florida, 2008.
[32] Seeingmachines, http://www.seeingmachines.com/ (October 2010).
[33] M. Nakayama, Y. Shimizu, Frequency analysis of task evoked pupillary response and eye-movement, in: Proceedings of the 2004 Symposium on Eye Tracking Research and Applications, ACM, San Antonio, Texas, 2004, pp. 71–76.
[34] K. Grauman, M. Betke, J. Gips, G.R. Bradski, Communication via eye blinks – detection and duration analysis in real time, in: Abstracts of the IEEE Computer Vision and Pattern Recognition Conference, vol. 2, 2001, pp. 1010–1017.
[35] F. Paas, J.E. Tuovinen, H. Tabbers, P.W.M. Van Gerven, Cognitive load measurement as a means to advance cognitive load theory, Educational Psychologist 38 (2003) 63–71.
[36] K. Cattapan-Ludewig, C.C. Hilti, S. Ludewing, F.X. Vollenweider, J. Feldon, Rapid visual information processing in schizophrenic patients: the impact of cognitive load and duration of stimulus presentation, Neuropsychobiology 52 (2005) 130–134.
[37] A. Laxmisan, F. Hakimzada, O.R. Sayan, R.A. Green, J. Zhang, V.L. Patel, The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care, International Journal of Medical Informatics 76 (2007) 801–811.

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

[38] I. Dror, A novel approach to minimize error in the medical domain: cognitive neuroscientific insights into training, Medical Teacher 33 (2011) 34–38.
[39] M.C. Beuscart-Zephir, J. Brender, R. Beuscart, I. Menager-Depriester, Cognitive evaluation: how to assess the usability of information technology in healthcare, Computer Methods and Programs in Biomedicine 54 (1997) 19–28.
[40] V. Aharonson, A.D. Korczyn, Human-computer interaction in the administration and analysis of neuropsychological tests, Computer Methods and Programs in Biomedicine 73 (2004) 43–53.
[41] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? Recognising natural interest by extensive audiovisual integration for real-life application, Image and Vision Computing 27 (12) (2009) 1760–1774.
[42] C.M. Schulz, E. Schneider, L. Fritz, J. Vockeroth, A. Hapfelmeier, M. Wasmaier, E.F. Kochs, G. Schneider, Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment, British Journal of Anaesthesia 106 (2011) 44–50.
[43] O. Palinko, A. Kun, Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies, in: Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, 2011.
[44] M. Causse, B. Pavard, J. Senard, J. Demonet, J. Pastor, Positive and negative emotion induction through avatars and its impact on reasoning performance: cardiovascular and pupillary correlates, StudiaPsychologica 54 (2012) 37–51.
[45] R.L.C. Mitchell, L.H. Phillips, The psychological, neurochemical and functional neuroanatomical mediators of the effects of positive and negative mood on executive functions, Neuropsychologia 45 (2007) 617–629.

Please cite this article in press as: S. Chen, J. Epps, Automatic classification of eye activity for cognitive load measurement with emotion interference, Comput. Methods Programs Biomed. (2012), http://dx.doi.org/10.1016/j.cmpb.2012.10.021

Attachment F

Multimodal Behavior and Interaction as Indicators of Cognitive Load

FANG CHEN, NATALIE RUIZ, ERIC CHOI, JULIEN EPPS, M. ASIF KHAWAJA, RONNIE TAIB, BO YIN, and YANG WANG, NICTA, Australia

High cognitive load arises from complex time and safety-critical tasks, for example, mapping out flight paths, monitoring traffic, or even managing nuclear reactors, causing stress, errors, and lowered performance. Over the last five years, our research has focused on using the multimodal interaction paradigm to detect fluctuations in cognitive load in user behavior during system interaction. Cognitive load variations have been found to impact interactive behavior: by monitoring variations in specific modal input features executed in tasks of varying complexity, we gain an understanding of the communicative changes that occur when cognitive load is high. So far, we have identified specific changes in: speech, namely acoustic, prosodic, and linguistic changes; interactive gesture; and digital pen input, both interactive and freeform. As ground-truth measurements, galvanic skin response, subjective, and performance ratings have been used to verify task complexity.

The data suggest that it is feasible to use features extracted from behavioral changes in multiple modal inputs as indices of cognitive load. The speech-based indicators of load, based on data collected from user studies in a variety of domains, have shown considerable promise. Scenarios include single-user and team-based tasks; think-aloud and interactive speech; and single-word, reading, and conversational speech, among others. Pen-based cognitive load indices have also been tested with some success, specifically with pen-gesture, handwriting, and freeform pen input, including diagraming. After examining some of the properties of these measurements, we present a multimodal fusion model, which is illustrated with quantitative examples from a case study.

The feasibility of employing user input and behavior patterns as indices of cognitive load is supported by experimental evidence. Moreover, symptomatic cues of cognitive load derived from user behavior such as acoustic speech signals, transcribed text, digital pen trajectories of handwriting, and shapes pen, can be supported by well-established theoretical frameworks, including O'Donnell and Eggemeier's workload measurement [1986] Sweller's Cognitive Load Theory [Chandler and Sweller 1991], and Baddeley's model of modal working memory [1992] as well as McKinstry et al.'s [2008] and Rosenbaum's [2005] action dynamics work. The benefit of using this approach to determine the user's cognitive load in real time is that the data can be collected implicitly that is, during day-to-day use of intelligent interactive systems, thus overcomes problems of intrusiveness and increases applicability in real-world environments, while adapting information selection and presentation in a dynamic computer interface with reference to load.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]—*Human information processing*; H.5.2 [**User Interfaces**]—*Interaction styles*

General Terms: Measurement, Experimentation, Human Factors

Additional Key Words and Phrases: Cognitive load, pen input, assessment, multimodal

The reviewing of this article was managed by associate editor Matthew Turk.
This work is supported by NICTA, which is funded by the Australian Government as represented by the Department of Broadband Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program. This research was also supported by the Asian Office of Aerospace Research and Development, grant no. FA2386-12-1-4049.
Authors' addresses: F. Chen, N. Ruiz, E. Choi, J. Epps (corresponding author), M. A. Khawaja, R. Taib (corresponding author), B. Yin, and Y. Wang, NICTA, Level 5, 13 Garden Street, Eveleigh, NSW 2015, Australia; email: {julien.epps, ronnie.taib}@nicta.com.au.

DOI 10.1145/2395123.2395127 http://doi.acm.org/10.1145/2395123.2395127

ACM Reference Format:
Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M. A., Taib, R., Yin, B., and Wang, Y. 2012. Multimodal behavior and interaction as indicators of cognitive load. ACM Trans. Interact. Intell. Syst. 2, 4, Article 22 (December 2012), 36 pages.
DOI = 10.1145/2395123.2395127 http://doi.acm.org/10.1145/2395123.2395127

1. INTRODUCTION

The past few decades have been marked by the rapid development of new and powerful information systems, granting access to volumes of data previously unheard of. The dramatic evolution of computers and networks has allowed exponential functionality to be offered by expert software. However, the capabilities of the human brain, working memory in particular, have remained unchanged and fairly limited. Even domain experts well-trained with the tools now struggle in the management of information. Worse still, the lack of metrics in complex environments makes it impossible to predict the tipping point, when the user no longer has control of the situation. This issue is exacerbated in high-intensity, data-laden, and safety-critical environments, and calls for a robust and real-time measurement of user's cognitive load.

Indeed, traffic management centers, crisis or air-traffic control rooms, and intelligent interactive systems involve inherently complex domain tasks for operators to solve. High cognitive demand arises from such tasks as mapping out flight paths, monitoring traffic, or even managing nuclear reactors. The ability to measure a user's cognitive load in real time can support personalized system adaptation to users affected by high cognitive load, easing the demand and avoiding stress, frustration, and errors. While conventional human-computer interaction paradigms (e.g., Graphical User Interfaces) are useful in personal computing applications such as word processing, they do not adequately support tasks that require the manipulation of complex data types and constraints in the way intelligent, interactive systems have the potential to do.

Our research goal for the past five years has been the development of technology for the implicit, objective, automated, and real-time estimation of a user's cognitive load, suitable for real-time deployment as part of an intelligent interactive system. The approach is focused on the identification of possible correlations between increasing levels of cognitive demand and both passive and active modalities, from speech, digital pen, and freehand gesture, to eye activity, galvanic skin response, and ElectroEncephaloGraphy (EEG). This article begins with a summary of the underlying psychological theories on which our research rests, an overview of our individual approach, and a review of the most promising indices and features that we have found to be sensitive to high cognitive load. Finally, we discuss the implications of our findings and plans for future work.

2. RELATED WORK

2.1. Working Memory and Cognitive Load

It is well-established that the two main limitations of working memory resources are its capacity and duration [Baddeley 1992]. According to Baddeley's model [1992], working memory has separate processors for visual and verbal information. Only a limited number of item "chunks" can be held in working memory at any one time and then only for a short amount of time [Cowan 2001]. These limitations are never more evident than when users undertake complex tasks or when in the process of learning, resulting in extremely high demands being placed on working memory. The construct of cognitive load refers to the working memory demand induced by a complex task in a particular instance where novel information or novel processing is required [Sweller et al. 1998]. Any single task can induce differing levels of mental effort or cognitive load from one user to another, or as a user gains expertise. This discrepancy in the mental

demand from person to person could be due to a number of reasons, for example, level of domain expertise or prior knowledge, interface familiarity, the user's age, or any mental or physical impediments. A task that may cause high load in one user may not necessarily do so in a more experienced user, for example.

The cognitive load construct comprises at least two separate load sources: *intrinsic load and extraneous load* [Paas et al. 2003; Sweller et al. 1998]. Intrinsic load refers to the inherent complexity of the task itself, whereas extraneous load refers to the representational complexity, that is, complexity that varies depending on the way the task is presented. In an intelligent interactive system, the inherent task complexity would be dictated by the domain. For example, in a traffic management scenario, a sample domain task may be to find the exact location of an accident. The equipment, tools, and applications the operator employs to complete the task, for example, a paper-based directory, a GIS, or electronic maps, or even street monitoring cameras, each contribute to extraneous load. Both of these types of load combine to form the overall experience of cognitive load. Situations that induce high levels of cognitive load can impede learning and efficient performance on designated tasks [Paas et al. 2003; Sweller et al. 1998].

The ability to determine exactly when a user is being cognitively loaded beyond a level that he or she is able to manage could enable the system to adapt its interaction strategy intelligently. For example, the system could attempt to reduce the cognitive load experienced by the operator—particularly in terms of extraneous load—such that optimal performance is facilitated. A number of methods have been used, both in Human-Computer Interaction (HCI) and other domains, to estimate the level of cognitive load experienced. There are four main methods comprising the state-of-the-art: subjective (self-report) measures, where users rank their experienced level of load on single or multiple rating scales [Gopher and Braune 1984]; physiological measures, such as galvanic skin response and heart rate [Delis et al. 2001]; performance measures, such as task completion time, speed, or correctness, critical errors and false starts [Gawron 2000; O'Donnell and Eggemeier 1986; Paas et al. 2008], as well as dual tasks [Chandler and Sweller 1991]; and finally, behavioral measures, which observe feature patterns of interactive behavior, such as linguistic or dialog patterns [Berthold and Jameson 1999], and even text input events and mouse-click events [Ikehara and Crosby 2005]. However, while most of these types of measures are suitable for research purposes, many are unfeasible for widespread deployment in interactive intelligent systems.

2.2. Subjective Measures

Traditionally, the most consistent results for cognitive load measurement have been achieved through subjective measures [O'Donnell and Eggemeier 1986]. These measures ask users to describe in fine detail and reflect each user's perception of cognitive load by means of introspection: the user is required to perform a self-assessment of his or her mental demand by answering a set of assessment questions immediately after the task. However, such an approach is impractical in real, day to day situations because the questionnaires not only interrupt task flow but also add more tasks to the load of potentially overloaded users.

2.3. Physiological Measures

The physiological approach for cognitive load measurement is based on the assumption that any changes in human cognitive functioning are reflected in human physiology [Kramer 1991]. The measures that have been used in the literature to show some relationship between subjects' mental workload or cognitive load and their physiological behavior include, among others, heart rate and heart rate variability [Kennedy and Scholey 2000; Mousavi et al. 1995; Nickel and Nachreiner 2000], brain activity (e.g., changes in oxygenation and blood volume, ElectroCardioGraphy (ECG),

ElectroEncephaloGraphy (EEG)) [Brunken et al. 2003; Wilson and Russell 2003], Galvanic Skin Response (GSR) or skin conductance [Jacobs et al. 1994; Shi et al. 2007], and eye activity (e.g., blink rate, eye movement, pupillary dilation) [Backs and Walrath, 1992; Iqbal et al. 2004; Lipp and Neumann 2004; Marshalle et al. 2003]. Changes in the physiological data occur with the level of stimulation experienced by the person and can represent various levels of mental processing. The data collected from body functions are useful as they are continuous and allow the signal to be measured at a high rate and in fine detail. However, physiological measures require users to wear a lot of cumbersome equipment, for example, EEG headsets that not only interfere with their task, but are prohibitive in cost and implementation. Additionally, the large amounts of physiological data that need to be collected and the expertise needed to interpret these signals render many types of physiological signals unsuitable for common interactive intelligent systems [Delis et al. 2001]. While they can be very sensitive to cognitive activity, the preceding issues in combination with the degree of variability of physiological signals, due to external factors such as temperature and movement, means they may have limited suitability for environments other than laboratory conditions [Delis et al. 2001].

2.4. Performance Measures

The hypothetical relationship between performance and workload as discussed by O'Donnell and Eggemeier [1986] is composed of three regions, A, B, and C as seen in Figure 1. The authors claim that primary task measures of workload cannot be used to reflect mental workload in Region Low, because this region is characterized as indicating "adequate task performance" on behalf of the subject [O'Donnell and Eggemeier 1986]. However, in many real-world tasks, what constitutes "adequate task performance" is analogous to a band of acceptable outcomes rather than a single correct response, and subtle differences may occur between different solution alternatives which may not be reflected in the overall performance measures used. As discussed in this article, we propose that certain features of the behavioral responses have the potential to differentiate between these solution outcomes by identifying compensatory behaviors. In much the same way, performance measures cannot measure spare capacity [Parasuraman et al. 2008], when a user still has plenty of cognitive resources to deploy. The relationship between cognitive load and performance, where it has been studied, is not as simple as might be hoped [Yeh and Wickens 1988]. Nevertheless, in this article we provide performance-based cognitive load classification results for experiments in which the task design and measurements made permit this.

In Region B, both primary and secondary task performance measures can be used to reflect workload as performance decreases. Dual-task approaches have been incorporated in several studies to measure subjects' performance in controlled conditions [O'Donnell and Eggemeier 1986]. While secondary task performance can provide a measure of remaining resources not being used by the primary task [Kerr 1973; Marcus et al. 1996], it is not feasible for operators to complete dual tasks "in the wild", and hence these cannot be adopted for widespread use. In real-world tasks, performance measures from the primary task can be extremely difficult to calculate on-the-fly, if at all. In the case of transport management Centers, senior staff will often conduct reviews of incident handling to debrief operators and qualitatively rate their performance. In this application, access to automatic cognitive load estimates around every hour or so would be considered a dramatic improvement for review purposes. Access every minute or so would be considered "real time" and could directly affect moment to moment operator allocation.

Performance measures tend to remain stable as load increases in Region B, particularly when the operator exerts a greater amount of mental effort, as noted by O'Donnell and Eggemeier [1986]. This is addressed more specifically by Hockey [2003]

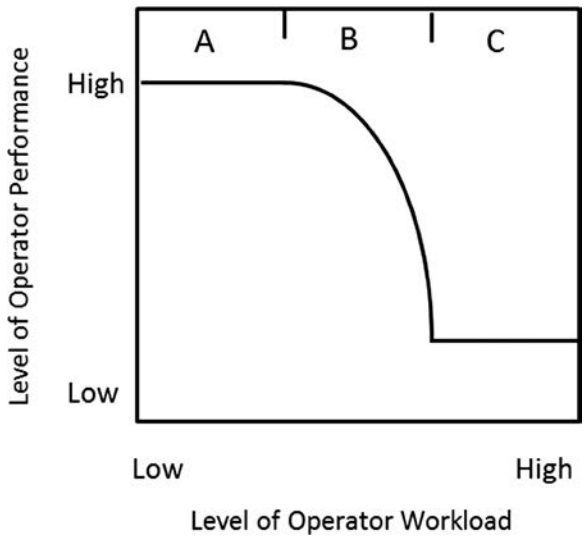


Fig. 1. Hypothetical relationship between workload and operator performance, adapted from O'Donnell and Eggemeier [1986].

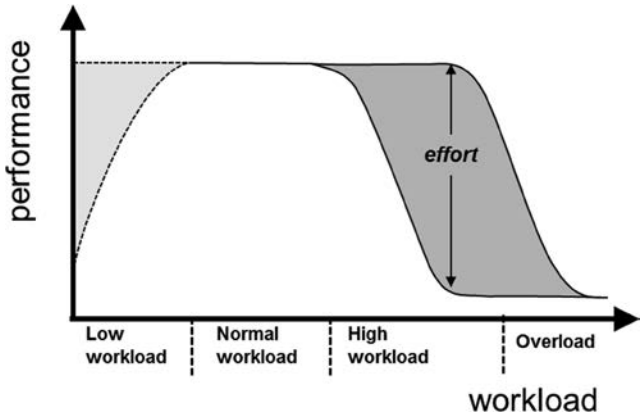


Fig. 2. Relationship between performance and workload, adapted from Hockey [2003].

who proposes a range within which compensatory efforts may have an effect. Figure 2 illustrates this concept; the subject still achieves a high level of performance within the region labeled "effort", depending on the degree of effort exerted. Exposure to high cognitive load (workload) culminates in a higher likelihood of errors [Byrne et al. 1998; Hockey 2003; Ruffell Smith 1979] and compensatory efforts can only be maintained for a time, after which the subject then fatigues and his or her performance begins to decline [Hockey 2003]. At the overload stage, compensatory efforts no longer make a difference; it is too late for the system to react appropriately to ease the operator's load and both the system and the user must engage in costly recovery strategies.

The approaches described thus far have the disadvantage of being physically or psychologically intrusive. Likewise, many of them are also post hoc and hence not conducive to the implementation of real-time adaptive behaviors and strategies by an interactive intelligent system or interface. Performance measures can also depend on the subject completing the task, which may not always be possible in high load

situations, for example, the subject may be stuck on one or two steps of the overall task for a relatively long period of time and no valid task-based performance assessment can be calculated.

Similarly, performance measures—which we define as measures that reflect the accuracy and correctness of a user's response and are directly relevant to the outcome of the task—are often calculated after the fact, if they can be assessed objectively at all. In the kinds of complex domains that we are targeting, measures based on performance outcomes are impossible to access in real time such that the system is able to act on the information in a timely manner. For example, the spontaneous nature of crisis management and other control-room situations means the user's performance in this sense is very difficult to rate, even during debriefing, and unique to almost every situation. The actions taken can vary widely from operator to operator, both in order and content, while still being equally effective in achieving the task goals and solving the problem to an adequate level of performance. In some cases, performance cannot be calculated automatically at all.

2.5. Behavioral Measures

On the other hand, we define response-based behavioral features as those that can be extracted from any user activity that is predominantly related to deliberate/voluntary task completion, for example, eye-gaze tracking, mouse pointing and clicking, keyboard usage, application usage, digital pen input, gesture input, or any other kind of interactive input used to issue system commands. These responses provide two types of information: first, the inherent content or meaning of the response, and second, the manner in which the response was made. For example, one could type in a sequence of numbers as part of a task in different ways using a variety of equipment, such as the keys on the top part of the keyboard (above the alphabet), or the keys on the number pad on the right side of the keyboard, or by clicking buttons on a numeric display with a mouse. The string of numbers is the same; this is the content or meaning in the response relevant to the domain task. The manner in which the response is made does not directly affect the outcome of the task, but does provide other information, for example, how long it took to enter the sequence of letters, how much pressure was exerted on each key, and in the case of the mouse usage, features such as the mouse trajectory and the time between clicks.

We define these sources as behavioral rather than performance-centric because the information they hold does not directly affect the domain-based outcome of the task, hence there is a lot of margin for differences within and between users. They are objective, and can be collected implicitly, that is, while the user is completing his or her task and without overt collection activities (e.g., stopping to ask the user to provide a subjective rating of difficulty), hence suitable for control-room-type environments. They are also distinct from physiological measures in that they are mostly or entirely under the user's voluntary control. Some of the measures we consider herein, for example, acoustic speech features, do not fall neatly into the usual definition of "behavioral", however, they share with behavioral measures the property of being nonintrusively and continuously acquired, they occur during a task rather than after it, and in most cases they are primarily under partial or full conscious user control, by contrast with post hoc measures (e.g., performance). While this is likely to introduce some variability relative to physiological measures, this may turn out to be smaller when the behavior is the response to a task or task type that occurs very often in the user's environment. There is evidence showing that these kinds of behavioral features can reflect mental states, such as mental effort and cognitive load. For instance, Gütl et al. [2005] used eye tracking to observe subjects' learning activities in real time by monitoring their eye movements for adaptive learning purposes. Although the visual functions are partly involuntary (e.g., the eye is drawn to salient items in the visual field), gaze is

under voluntary control, and can be considered a behavioral measure. Contemporary eye trackers do not require cumbersome headsets and can be extracted from video collected through standard Webcams. Others have used mouse clicking and keyboard key-pressing behavior to make inferences about their emotional state and adapt the system's response accordingly [Ark et al. 1999; Liu et al. 2003].

Previous research also suggests the existence of major speech cues that are related to high cognitive load [Berthold and Jameson 1999; Jameson et al. 2009; Keränen et al. 2004; Müller et al. 2001]. Examples of features that have been shown to vary according to task difficulty include pitch, prosody, speech rate, speech energy, and fundamental speech frequency. Some studies have reported an increase in the subjects' rate of speech as well as speech energy, amplitude, and variability under high load conditions [Brenner et al. 1985; Lively et al. 1993]. Others have found specific peak intonation [Kettebekov 2004] and pitch range patterns [Lively et al. 1993; Wood et al. 2004] in high load conditions. Pitch variability has also been shown to potentially correlate to cognitive load [Brenner et al. 1985; Tolkmitt and Scherer 1986; Wood et al. 2004]. These features are classified as behavioral because they show variations regardless of the meaning of the utterance being conveyed.

Higher-level-features, such as linguistic and grammatical features, may also be extracted from user's spoken language for patterns that may be indicative of high cognitive load. Significant variations in levels of spoken disfluency, articulation rate, and filler and pause rates [Berthold and Jameson 1999] have been found in users experiencing low versus high cognitive load. Extensions of this work attempt to recognize cognitive load levels using a Bayesian network approach [Müller et al. 2001]; other work has found changes in word frequency and first-person plurals [Sexton and Helmreich 2000]. Changes in linguistic and grammatical features have also been used for purposes other than cognitive load measurement [Schilperoord 2001].

More closely linking multimodal interactive systems and cognitive load, users have been found to change and adapt their multimodal behavior in complex situations. Empirical evidence suggests that when tasks are more difficult, users prefer to interact multimodally rather than unimodally across a variety of different application domains [Oviatt et al. 2004]. As task complexity increases, users tend to spread information acquisition and production over distinct modalities, seemingly for more effective use of the various modality-based working memory resources [Alibali et al. 2000; Goldin-Meadow et al. 2001; Mousavi et al. 1995; Oviatt 1997, 2006]. Temporal relationships that exist between interaction modalities (e.g., speech and pen) have also been shown to change under increased load conditions, showing a deeper entrenchment into the participant's preferred multimodal pattern, either simultaneous or sequential [Oviatt et al. 2004]. Another study employed users' digital-pen gestures and usage patterns to evaluate the usability and complexity of different interfaces [Oviatt 2006]. It has been suggested that pen-based interfaces can dramatically improve subjects' ability to express themselves over traditional interfaces because linguistic, alphanumeric, and spatial representations bear little cognitive overhead [Oviatt 2009].

2.6. Estimating Load from Interactive Behavior

The premise of our research is that observations of interactive features may be suitable for cognitive load assessment because a user experiencing a high cognitive load will show behavioral symptoms relating to the management of that load. This suggests a more generalized effect of an attempt to maximize working memory resources during completion of complex tasks [Mousavi et al. 1995; Oviatt et al. 2004]. High cognitive load tasks increase the cognitive demand, forcing more cognitive processes to share fewer resources. We hypothesize that such reactions will cause changes in interactive and communicative behavior, whether voluntary or otherwise.

The hypothesis that behavioral responses can provide insight into mental states and processing is not without precedent. Spivey et al. contend that reaching movements made with a computer mouse provide a continuous two-dimensional index of which regions of a scene are influencing or guiding "action plans", and therefore reflective of changes in cognitive processes [Spivey et al. 2005]. In an experiment involving decision making, McKinstry et al. found that mouse trajectories for answer selection (YES and NO) options are characterized by the greatest curvature and the lowest peak velocity when the "correct" choice to be made is more ambiguous or more complex [McKinstry et al. 2008]. They conclude that spatial extent and temporal dynamics of motor movements can provide insight into high-level cognition [McKinstry et al. 2008; Rosenbaum 2005]. Dale et al. ran a study where participants' hand movements were continuously tracked using a Nintendo® Wii^{TM} remote, as they learned to categorize elements [Dale et al. 2008]. They noted that participants' arm movements started and finished more quickly and more smoothly (decreased fluctuation and increased perturbation) after learning the categorization rules. The "features of action dynamics" show that participants grow more "confident" over a learning task, and indicate learning has taken place. Galen and Huygevoort have shown that time pressure and dual-task load results in "biomechanical adaptations of pen pressure" as a coping mechanism to increased load [Galen and van Huygevoort 2000]. These studies provide evidence that features of behavioral responses can be harnessed to provide an indication of changes in cognitive processing and strategy.

Symptomatic changes in structure, form, and quality of communicative and interactive responses are more likely to appear as people are increasingly loaded, as will be described in this article. With the proliferation of sensor data that can be collected from users through the latest intelligent systems, there is a very specific opportunity to use this behavioral input to detect patterns of change that are correlated with high load, and use these cues to guide the adaptation strategies employed by the system. Here, "patterns of change" is the term we use to describe any behavioral change, while we consider that cues are perceptible or observable behaviors that can be used to signal that a change is occurring. Such features have the added advantage of offering an implicit (as opposed to overt) way to collect and assess cues that indicate changes in cognitive load. However, to do this it is necessary to first identify and quantify the fluctuations of features in user interaction as cognitive load varies over a variety of input modalities.

The major challenge of choosing the assessment features for automated load detection is to make sure they satisfy the requirements of consistency, compact representation, and automatic acquisition [Yin et al. 2008]. Our aim is to find effective features that reliably reflect the cues and can be extracted automatically such that they are useful in adaptive systems. A second goal is to find a suitable learning or modeling scheme for each index to resolve the corresponding level of cognitive load [Yin et al. 2008]. By manipulating the level of task complexity and cognitive load, and conducting a series of repeated-measures user studies in a variety of scenarios, we have been able to identify a series of cognitive load indices based on features from a number of input modalities, specifically, observations of significant changes in speech and digital-pen input that are abstracted from individual application domains in which they occur as well as correlated to high cognitive load. In this work, we use the term "indices" to denote operationalized cues that can be resolved by a machine, and may comprise many individual features (which may or may not be indicative of load on their own).

3. SPEECH-BASED FEATURES OF COGNITIVE LOAD

Speech signal features can be a particularly good choice for cognitive load indices, since speech data exists in many real-life tasks (e.g., telephone conversations, voice

command and control systems, self-talk) and can be easily collected in a nonintrusive and inexpensive way with a close talk microphone. Types of speech features can vary, from intensity, pitch, and formants inspired by speech production, to other acoustic, prosodic, or linguistic features such as grammar and syntax. We have explored all types of features with significant results.

3.1. Speech Datasets

Over the last five years, we have conducted a series of user studies in which we have collected a large amount of interactive and conversational speech data in a variety of application domains, ranging from speech responses to simple psychological tests such as the Stroop test [Stroop 1935], to reading and comprehension speech, to interactive speech (in both simulated and real multimodal interactive system environments) and think-aloud speech from controlled user studies with interactive systems [Le et al. 2010a, 2010b; Stroop 1935; Yap 2011; Yap et al. 2010a, 2010b; Yin and Chen 2007; Yin et al. 2008]. All data was collected through a series of specially designed and controlled experiments, where we have manipulated multiple parameters to isolate different cognitive load factors. Finally, through collaborative partnerships with industry, we have also collected speech from the field, generated in real-life environments such as air traffic control rooms, call centers, and bushfire control training exercises.

The speech datasets we have used in our investigations have either been elicited during tasks of increasing cognitive load a priori or labeled a posteriori with expert ratings of task complexity and cognitive load [Yin et al. 2008, 2007]. While six different speech datasets have been collected in our team—including field data, and lab studies featuring multimodal interaction with speech and gesture, multimodal interaction with speech and pen in two different application domains (incident management and basketball training), as well as a simulated driving user study—two key databases have been used in the development of the speech cognitive load measurement system. The first is the Stroop database and the second is the Reading and Comprehension database. Both were generated from lab controlled experiments featuring cognitive load manipulations.

The Stroop test corpus is based on the original test by John Ridley Stroop [1935]. Three levels of cognitive load were derived from this paradigm, the task difficulty arises from cognitive interference when reading color names or naming color words. In our version of the Stroop test, speech from the low cognitive load task was recorded by asking the subjects to read aloud a series of words (which were the names of colors) written either in black font or a congruent font color (e.g., the word red, written in red font). During the medium load level, subjects were asked to name the font color of words written in incongruent color (i.e., the font color of the words is different from the meaning of the word, e.g., white written in blue font). In the high cognitive load level, a time constraint was added to the medium load task, forcing the subjects to complete the task faster. An additional recording was collected from each subject when asked to read a short story aloud for approximately 90s; this was used as baseline data and for the background model of the base cognitive load measurement engine illustrated in Figure 3. This corpus contains single-word utterances of ten color names, spoken slowly in a series. The majority are also very short, containing only one or two syllables (e.g., red, blue, white). In addition, there is a speech rate artifact caused by the time constraint for the high cognitive load speech.

The second corpus, the Reading and Comprehension corpus, was generated by asking the subjects to read three stories aloud, each corresponding to a load level from low to high. The difficulty levels of the stories were estimated based on the Lexile scale [Lennon and Burdick 2004], a semantic difficulty and syntactic complexity measure scale ranging from 200 to 1700 Lexiles (L), corresponding to the reading level expected

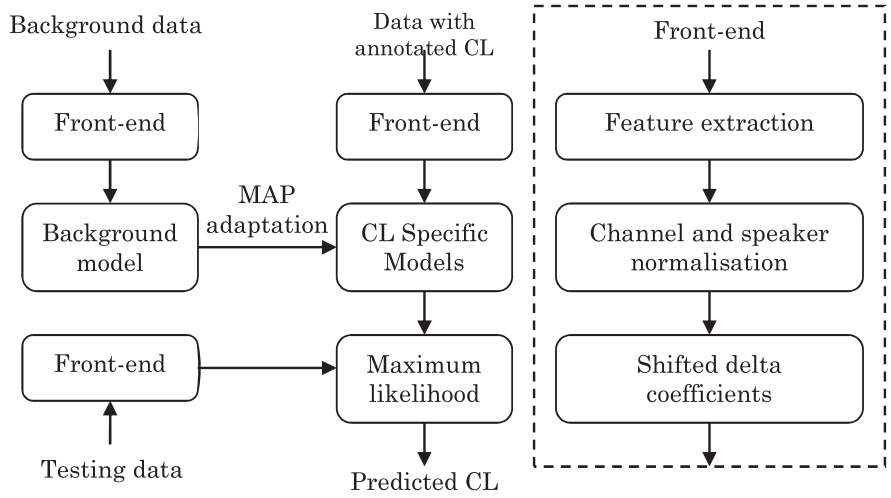


Fig. 3. The base system structure for acoustic and prosodic CLM.

from a first grade student to a graduate student. The Lexile ratings of the stories used were 925 L, 1200 L, and 1350 L, respectively. The approximate lengths of the utterances corresponding to reading the story for the low, medium, and high cognitive loads are 90s, 140s, and 230s, respectively. The story reading speech is referred to as the Reading data. After each story, the subjects were asked to answer three open-ended questions related to the content of the story. The approximate length of each answer to the three questions for all three levels of cognitive load is 30s. In contrast to the Stroop dataset, the Reading and Comprehension corpus contains a significantly larger vocabulary due to the less-constrained content of the stories and the answers given to the questions.

3.2. Acoustic and Prosodic Speech Features

Inspired by previous research on emotional and stressed speech [Fernandez and Picard 2003; Hansen 1996; Picard 1997], we expected that prosodic patterns (i.e., voice pitch variation) could be used as a cue to reflect cognitive load. The rate of pauses and rate of pitch peaks emerged as good potential indicators of cognitive load levels for the speech in multimodal interaction tasks in a study described in Ruiz et al. [2006]. We used a sliding window implementation, which showed these indicators to be higher when the cognitive load level was higher [Yin and Chen 2007]. This proved to be the first of the speech-signal-based indices we uncovered.

Since the areas of interest in cognitive load monitoring are extreme levels of cognitive load (too high or too low), the assessment problem was reinterpreted from a continuous scale of degrees of cognitive load by introducing the notion of discrete levels of load. A classification approach could then be employed for cognitive load measurement [Yin et al. 2008, 2007]. In a bottom-up, data-driven strategy for cognitive load assessment, the subsequent datasets were employed in a statistical machine learning approach, in an effort to build a cognitive load monitoring engine based solely on changes in the speech signal [Yin et al. 2008, 2007]. We have developed an automatic, real-time, speaker-independent cognitive load assessment module that can be adapted to varied task scenarios [Yin et al. 2008, 2007].

A Gaussian-Mixture-Model (GMM)-based classifier [Reynolds and Rose 1992], forming part of the base system as pictured in Figure 3, was created with semisupervised training, from hours of annotated data from both of these sets, where each of

the cognitive load levels is modeled by a GMM. The classification engine determines the best-matched model based on a calculated likelihood score. Channel and speaker normalization are deployed also for improving robustness. The classification process uses a mixture of frame-based acoustic features: Mel-Frequency Cepstral coefficients (MFCCs), pitch, and intensity. MFCCs are a set of features commonly used in speech recognition applications, and they capture information in the magnitude part of the speech spectrum. Pitch and intensity, on the other hand, are features that capture information relating to the prosody of speech. Additionally, a background model was introduced, in the form of another GMM trained on data from all the cognitive load levels. All individual cognitive-load-level models are adapted from it using the Maximum A Posteriori (MAP) estimation technique. Since the background model models the basic feature distribution shared by all speakers under all load levels, it is a good initial distribution from which to adapt models of specific levels, and therefore improves the generalization capabilities of models of specific CL levels when training and/or test data are limited.

The classification accuracy for both of these databases using the baseline MFCC cognitive load measurement system has been very positive. The Stroop test data reveals an accuracy of 78.9% for classification into low, medium, and high load in a speaker-independent scenario [Yin et al. 2008]. Similarly, over three discrete cognitive-level ranges, classification of the Reading and Comprehension dataset (comprehension data) achieved a 71.1% accuracy in a speaker-independent closed-set setting [Yin et al. 2007].

MFCCs proved to be an effective set of baseline frame-based features for cognitive load classification. However, MFCCs do not provide us with any insight into how cognitive load affects the speech spectrum or the underlying speech production system. Moreover, since MFCCs may have higher dimensionality than is strictly required for the problem, it may be possible to achieve the same result using more targeted sets of features. Voice source or glottal features have been investigated in an attempt to link cognitive load to the speech production system [Le et al. 2010; Yap et al. 2010a] with some success. The system tested on the Stroop test corpus, after fusing the scores of the baseline system (combination of MFCC, pitch, and intensity) with the scores of a glottal-parameter-feature-based system [Yap et al. 2010c] produced an accuracy of 84.4% on that dataset. By way of comparison, a performance measure (the correctness of the answer) yielded 57.1% accuracy when classified across the three load levels. Furthermore, the difference in answer correctness across medium and high load levels was not statistically significant ($p = 0.265$).

Investigations conducted in order to assess the effectiveness of detailed spectral features such as Spectral Centroid Frequency (SCF) [Paliwal 1998] and Spectral Centroid Amplitude (SCA) [Le et al. 2011], as part of the cognitive load classification system, have also proven successful. Inclusion of these features has resulted in improvements to the baseline classification result. First, the Stroop test classification over three levels reaches 88.5% accuracy with the fusion of the SCF-based system and the SCA-based system. Second, the speaker-dependent system based on the fusion of the SCF-based system and the SCA-based system has an accuracy of 84.3% in the Reading and Comprehension corpus.

A recent step has been to study the effect of cognitive load on the vocal tract through an investigation of formant frequencies. Formant frequencies (the frequencies at which broad spectral peaks occur in the magnitude spectrum of speech) are closely related to the underlying configuration of the vocal tract. The results show that 2-class (low and high) and 3-class (low, medium, and high) utterance-based evaluations on both of the databases, using frame-based formant features, perform at least as well as the baseline system with MFCC features. This is despite formant features having a dimensionality of 3 compared with MFCCs with a dimensionality of 7 [Yap et al. 2011].

This finding suggests that cognitive load information can be captured using features with lower dimensionality [Yap et al. 2011], potentially reducing the amount of data needed for training models. Combinations of features derived from formants and speech production models have produced accuracies to 95% in more recent work [Yap 2011].

It is possible to conduct both linguistic and acoustic analysis on any speech data, including speech from teams. However, suitable microphones are needed (individual close talk microphones are preferable) and more involved preprocessing may be required in group situations for acoustic analysis, since there is a higher chance of crosstalk noise in the recorded speech signals.

3.3. Real-Life Case Studies[1]

As part of our research, the involvement of collaborative industry partners has been sought in order to obtain field data on which to test our behavioral indices, in particular, the speech features. The following case studies show the viability of using speech-based cognitive load indices for measurement and assessment with our system.

The first case study took place in an Emergency Communications Center in North America, where the operators are responsible for receiving information from police, traffic authorities, and other sources, and dispatching them to relevant ambulance units. A total of 37 working sessions were recorded from 10 participants during training. Each 30 minute session contained a total of 1113 events under three workload levels, each lasting approximately 10s (i.e., each cognitive load estimate was made after 10s). All events were manually annotated with an observed workload level during a postreview session by domain experts to label the data for adaptation and evaluation purposes. There were 599 low load events, 465 medium load, and 49 high load events in the dataset. Our speech-based cognitive load measurement system successfully estimated the load level of participants with an average accuracy of 82.2% over 3 levels of load when it was evaluated with a 10-fold leave-one-participant-out cross-validation. Moreover, the high load event detection achieved a 95.9% hit rate and 4.1% false alarm rate.

In a second case study conducted at a large outsourced Contact Center operator in Australia (5000+ seats), high personnel attrition rates and associated hiring and training expenses were key issues to be addressed. Our speech-based cognitive load measure was used to investigate the correlation between tenure and demonstrated load level under a series of tests. By analyzing the speech responses of the potential candidates, it was possible to predict whether the candidate was likely to perform well as a contact center operator, and hence was more likely to have a longer tenure. A group of 191 freshly hired agents received the assessment, and the attrition reduction was evaluated over 12 weeks. The overall attrition rate reached 18% at the end of week 12, while the attrition rate of the most suitable candidates, as identified by the system, was only 9%, representing a relative 50% improvement over existing assessments. Cognitive load estimates were made once every 2–3 minutes of speech recorded. For this application there was no requirement for real-time processing, however, a demonstrator system for call center monitoring developed by the authors yielded a reliable cognitive load estimate every 3s (less than a typical utterance).

A third case study conducted on real-life training data involved air traffic controllers from 3 different regional airports in Australia. The speech data produced by the controllers in their communications with the pilots during a shift was recorded. In addition, every two minutes, the controllers were also asked to report on their current level of cognitive load. The tasks were designed to emulate different difficulty levels, for example, from routine landings of single flights to multiple landings in inclement weather. The evaluation dataset consists of speech data collected from 10 controllers

[1] Industry partners are kept anonymous for confidentiality purposes.

over a total of 26 sessions (about 3 sessions per controller). Each session contains 15 speech segments and each segment has a fixed duration of 2 minutes. Since there may be pauses or silences in a segment, the actual speech duration is only about 40s on average. For each controller, half of the available segments within a session were used for the training of models, while the other half were used for testing. The annotation of the individual segments was obtained by mapping the corresponding post hoc subjective ratings into 4 cognitive load levels: low, medium, high, and extremely high. Our speech-based cognitive load measurement was applied on the evaluation dataset, resulting in an average 87.9% accuracy rate in the identification of the cognitive load level. Cognitive load estimates were made once every 2 minutes of speech recorded.

3.4. Linguistic Features

In a complementary, semantically-driven approach, we have also examined the linguistic features of speech for cues that indicate high cognitive load. It has been shown that people's selection of the language elements and linguistic features varies from one situation to another depending on the circumstances of the situation [Dechert and Raupach 1980; Sexton and Helmreich 2000]. We have been successful in isolating a number of cognitive indices based on linguistic features that correlate strongly with high load. The data examined here have been gathered from one of three (and in many cases, more than one) scenarios: Reading and Comprehension lab study; the Bushfire training team lab study; or the real-life Bushfire team training field study. The linguistic features of interest are: pause features, grammar features, language complexity features, and word features.

Pause Features. Traditionally in psychology, the pauses during natural speech have been associated with a person's thinking and cognitive processes. It is argued that the more time it takes to produce the response, the more cognitive energy it requires to do so [Schilperoord 2001]. In other words, the increased amount of time spent in pausing (and hence thinking) while talking represents the increased level of cognitive load experienced [Esposito et al. 2007; Schilperoord 2001]. We have found that people use more and longer pauses (including both silent and filled pauses) under high cognitive load conditions versus low load conditions. Furthermore, it was found that people's response latency increases, confirming results from other studies in the literature [Berthold and Jameson 1999; Müller et al. 2001].

Grammatical Features. The use of personal pronouns has also been found to differ significantly, specifically the use of individual and collective pronoun use in team-based tasks. Four personal pronoun words (1^{st}-person singular, 1^{st}-person plural, 3^{rd}-person singular, and 3^{rd}-person plural) were examined in low versus high load conditions. The results show an interaction between usage of pronoun types (singular versus plural) and task difficulty (and so the cognitive load). People's use of singular pronouns decreased while their use of plural pronouns increased significantly when cognitive load was high. As task difficulty increases, teams tend to share more of the work [Kirschner et al. 2009] and this behavior is visible through their pronominal usage preferences. A further analysis of the results confirmed that use of both singular personal pronoun words (1^{st}-person and 3^{rd}-person) decreased while their use of both plural pronoun words (1^{st}-person and 3^{rd}-person) increased when cognitive load was increased.

Language Complexity. The complexity of a written or spoken text or transcript can be measured by two main factors: semantic difficulty and syntactic complexity [Lennon and Burdick 2004]. Our investigations show that while working collaboratively and performing tasks of high difficulty, people speak more and use longer sentences as the cognitive load increases. That is because under high workload conditions, as things become more complex, team members communicate more and provide more explanations as a strategy to deal with high task difficulty [Katz et al. 1998]. The language complexity

measures we used include lexical density [Chalker and Weiner 1998; Ure 1971], complex word ratio [Chalker and Weiner 1998], Gunning Fog Index [Gunning 1952; Reck and Reck 2007], Flesch-Kincaid Grade [Flesch 1948], SMOG Grade [McLaughlin 1969], and Lexile Level [Lennon and Burdick 2004]. While these complexity measures have mostly been used for written texts, for example, articles and essays, we have successfully demonstrated their use for measuring people's cognitive load from their spoken or written texts [Khawaja et al. 2010]. People's lexical density, that is, their use of unique and different words (or vocabulary richness), decreases as cognitive load increases: a result of fewer working memory resources available for the language processing task [Baddeley 2003], resulting in poorer selection of many unique words from the pool of words stored in long-term memory. A second result reveals people's spoken language becomes more complex and difficult to comprehend under high load conditions (again, working memory resources are allocated to the task itself rather than on speaking) and as a result, their speech is complicated, and comprised of often ill-formed sentences.

Word Categories and Valence. Qualitative investigations into word category usage show that while performing a collaborative task, as the cognitive load increases, people's use of negative emotion words significantly increases and their use of positive emotion words decreases significantly. The analyses also show that people used on the average fewer overall emotion words (either negative or positive) under high load situations as compared to low cognitive load. More importantly we found a significant interaction between the emotion words (negative versus positive) and cognitive load levels (low versus high). People working in groups and exhibiting negative emotions spend more time negotiating and engaging in more group discussions [Donner and Hancock 2011], which also supports our previous findings of increased word count in high load. Members of teams experiencing high load used significantly more cognitive mechanism words (e.g., think, consider, know, remember) and perceptual words (e.g., hear, view, touch) than those under low load conditions portraying their increased mental effort and concentration for the task. Similarly, their use of conflicting and disagreement words (e.g., no, wrong, never) increased significantly while there was a significant decrease in the use of agreement words (e.g., ok, right, fine). Other studies show people experiencing negative emotions tend show more disagree ment [Hancock et al. 2008].

The linguistic and grammatical features can be used as an individual set of cognitive load indices in the domains where speech and/or conversational transcripts are used as the main input modalities. Cognitive load measurement from the proposed linguistic features will require state-of-the-art Automatic Speech Recognition (ASR) technology with highly accurate automatic Speech-To-Text (STT) functionality to realize its potential in practical applications. Our linguistic approach to measuring cognitive load may also be used as a post hoc analysis technique for user interface evaluation and interaction design improvement, in addition to the acoustic speech analysis of load.

4. DIGITAL PEN-BASED FEATURES OF COGNITIVE LOAD

Digital pen or digital ink is becoming an increasingly popular method for interaction in specialized systems, beyond use as a pointing device. Recognition accuracy has markedly improved for handwriting purposes, symbolic drawing and gesture recognition, and sketching and visualization, as well as mark-up/annotation applications. An intuitive input mechanism, pen usage is said to be spontaneous, allowing users to produce two-dimensional representations almost as quickly as they are envisaged [Oviatt 2009; Schwartz and Heiser 2006]. In particular, pen input can support thought-organizing activities such as counting, ordering, grouping, labeling, and showing relationships, helping complex problem solving in high load contexts [Oviatt 2006], and is thus an ideal candidate for capturing symptomatic input changes. Geometric and temporal features of ink trajectory can be used as potential cues suited for automated

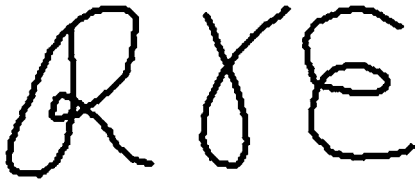


Fig. 4. Pen gestures: traffic flow, distance cost, and toggle route.

extraction, while higher-level recognition of characters and meaning can be compiled in a post hoc manner to offer an alternative view of how pen input is affected by high load tasks. Our investigation of cognitive load assessment features from digital pen spans three types of input: pen gesture input of predefined shapes, handwriting and freeform pen input, including note taking and sketching.

4.1. Symbolic Pen Gesture Features

Symbolic gestures refer to pen input methods that require the user to reproduce a specific two-dimensional shape modeled on a predefined shape to trigger a specific function within an interactive intelligent system. Our motivation for examining pen-input features when cognitive load is high is based on the premise that a user's performance is likely to be affected at a fine-grained level, where the quality of his or her motor productions may diminish in much the same way as his or her speech signal, due to low working memory resources [Ruiz 2011]. In fact, empirical evidence we have collected shows that the degeneration geometric features in predefined pen-gesture shapes increases significantly when cognitive load is very high, suggesting that a cognitive load index could be derived from such a measure [Ruiz et al. 2007].

Pen-gesture ink trajectories used in our analysis were collected using a custom interactive application, where users were required to build alternative routes on a map. The cognitive load factor was manipulated through three levels of task complexity, requiring users satisfy increasing sets of constraints related to the distance and traffic congestion of roads along possible alternative routes [Ruiz et al. 2007]. We defined three types of predefined functional pen-gesture inputs: to query the traffic flow, distance cost, and a toggle function showing the start and end of the route being constructed. These were invoked when the user drew any of the possible pen inputs, shown in Figure 4, on the map area. It was expected that when cognitive load increases, the shapes produced as part of the user's input would degrade in quality, that is, differ more significantly from the standard form of that shape. Dissimilarity could be due to asymmetry, jittery strokes, or generally "messy" script. Users produced a set of pen-gesture inputs in a "no-load" task (essentially 10 instances of each type of gesture on a blank screen), in order to create a standard form for that user from these trajectory instances.

The Mahalanobis distance (MDIST) was used to measure the level of degeneration of each single-stroke shape instance to the standard form [Rubine 1991; Ruiz 2011]. MDIST is most often used for recognition purposes, to classify the sample input into its correct type, however, we were able to leverage this function by also recording the degree of difference between the baseline form of each pen-shape type for each user and comparing to standard features. MDIST is a statistical measure of how similar an unknown vector of features is to a known vector of features, and considers the correlations between the features being assessed [Rubine 1991]. This is done via a classification technique that calculates MDIST as the weighted Euclidean distance from the vector of sample points (each input trajectory) to a standard model created for the inputs of that type, for each user. Subsequent pen inputs, produced during task time, were generated in each load level and MDIST was used to quantify the degree of

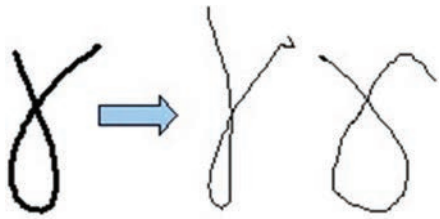


Fig. 5. Standard form and degenerate samples.

dissimilarity between the standard form and each specific instance. Hence, the greater the distance from its own type, the higher the MDIST value and geometric difference between the sample and the standard. The changes in MDIST have been found to be statistically significantly different between low cognitive load and high cognitive load tasks [Ruiz 2011; Ruiz et al. 2007].

Some examples of a standard form and sample inputs are shown in Figure 5. A MDIST result of zero is interpreted as a perfect replicate of the standard form of the shape for that user. The set of MDIST values generated by the system were developed by using a variant Rubine's specification for a classification-based gesture recognizer [Rubine 1991].

Using MDIST as a proxy for the geometric degeneration of pen shapes has a number of useful properties. Firstly, it can combine a number of individual geometric features into a single-value measurement of degeneration. This means that the trajectory changes from the baseline, or standard form, can occur in any or all of the features used by the classifier. MDIST can combine very small geometric changes in multiple features to register a significant combined change from the standard, and this type of deformation can be compared with large geometric change in a single feature. In this sense, it is a "true" measure of degeneration of the shape, with respect to any of the features used. Similarly, the same set of geometric features can be used uniformly to measure degeneration of all types of single-stroke pen shapes, hence providing comparable results for a wide variety of shapes and allowing us to group the data to obtain higher level of confidence in the assessment of degeneration.

In high load situations, a combination of both high intrinsic and high extraneous load can negatively affect both intrinsic and extraneous types of processing. This is reflected in decreased performance scores (intrinsic) as well as degradation in the quality of modal productions (extraneous). The degree of degradation as measured by MDIST can provide an indication of increased load between extreme load levels (low and high) for 85% of subjects, regardless of expertise. This measure is also personalized; each user has his or her own standard form that can be updated as often as necessary, and the baseline can be updated over time, so that the user is not penalized for learning to use the pen input more efficiently over time, for example, holding the pen differently or more comfortably [Ruiz 2011].

4.2. Handwriting Features

While the technical challenge of handwriting recognition has received ample attention from HCI researchers, analysis of form and structure of handwriting itself has received relatively little, and certainly not in the context of cognitive load assessment. Research into cognitive load during handwriting is important for improving the performance and experience of users in pen-based interactions [Frankish et al. 1995]. Our team's investigations have resulted in the first pen-based cognitive load classification engine for handwriting.

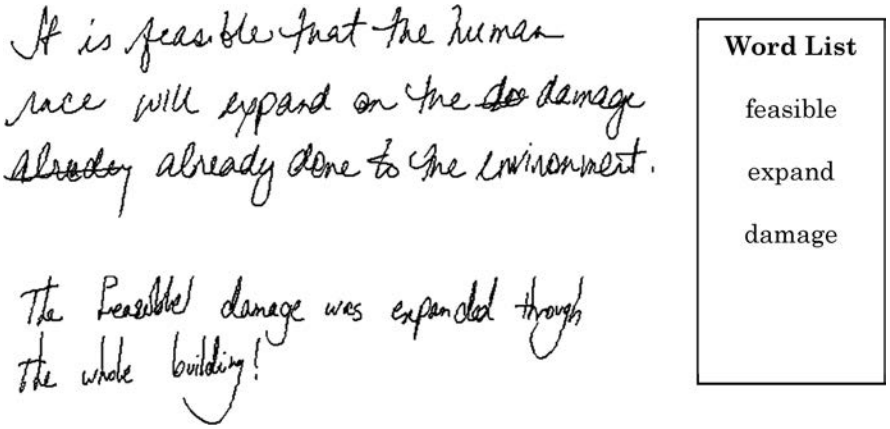


Fig. 6. Handwriting samples for high load task from two different users [Yu et al. 2011c].

According to handwriting experts, the writing process comprises three distinct phases: planning, translating, and reviewing [Vanderberg and Swanson, 2007]. The cognitive load induced by these processes places a demand on working memory resources from each of these in subsequent stages. We hypothesized in the same way as low-level trace features were found to change in pen-gesture inputs, stroke-level changes may also be detectable in handwriting produced under high load. An initial attempt to statistically analyze the stroke-level features of velocity, length, and pressure information with respect to increasing cognitive load showed that local maximum writing pressure, and local minimum writing velocity for strokes in particular, are sensitive to the cognitive load of the writer.

The findings are based on a study where subjects completed a sentence composition task using sets of given words. The number of words that were to be used in each sentence increased in each of the three complexity levels. Subjective ratings confirm the levels induced appropriate cognitive load differences. The handwriting dataset produced included approximately 600 handwritten sentences. Sentences from two subjects are shown in Figure 6 as examples.

The analysis shows that maximum writing pressure tends to occur at the beginning, at the corners, and at the end of strokes, where the minimum writing velocity is observed concurrently [Yu et al. 2011a]. This could be attributed to the shaping of alphanumeric letters, as it appears that writers experience higher cognitive load when forming the shapes than when producing straight parts of a stroke [Yu et al. 2011a]. Straight sections do not require a change in the direction of the pen trace, allowing the writer spare resources for other cognitive processes, such as reviewing of previously written material. This may suggest that cognitive load can potentially fluctuate even during the process of a stroke, correlated with the tempo of stroke construction [Yu et al. 2011a].

A second attempt has been made to use sample-based rather than stroke-based features from the ink trace [Yu et al. 2011b]. Specifically, this meant examining each writing point as a set of attributes including time-stamped trajectory coordinates and pressure of the pen tip, and the orientation of the pen tip [Yu et al. 2011b]. This information was modeled using Gaussian mixture models. Taking the combination of pressure and azimuth as features of the pen trace and using the same classifier, the application of altitude intervals improved the classification accuracy from 50.1% to 63.5%. Also, a particular span of pen altitudes, corresponding to about 12% of the writing samples, was found to produce a higher cognitive load classification accuracy of 75.4%. Using

Fig. 7. Sample digital notepad input.

altitude to sort the samples used in the models resulted in significant improvement, which signified that for samples with similar altitude within the low altitude interval, their pressure and azimuth attributes are sensitive to cognitive load changes. This finding could potentially decrease the computational cost for pen-based cognitive load classifications [Yu et al. 2011b]. For purposes of comparison, a performance-based measure (the rate of modification/correction after each writing task) yielded a classification result of 47.4% using a Parzen window classifier.

As a nonintrusive supporting component, a cognitive load measurement module based on handwriting with a digital pen can provide a useful reference to control the difficulty level of tasks where a writing requirement exists. In order to use the engine, a user would initially need to set up a profile to log the individual characteristics of his or her handwriting (e.g., altitude distribution) by producing a sufficient amount of text [Yu et al. 2011b]. The quantity of text needed is a trade-off between profiling time and system accuracy. [Yu et al. 2011b].

4.3. Freeform Pen Features

The use of digital pen for freeform note-taking, including sketches, symbolic diagrams, and other miscellaneous "doodles", is less common in high-pressure control-room environments, despite the fact that many operators use an array of low-tech tools, such as physical paper and pen to support their work processes. Indeed this has been found in a variety of domains [Lajoie 2000; Schwartz and Heiser 2006]. In our industry case studies, in particular with a large traffic monitoring center, we have seen that some of the work processes are duplicated, with operators transferring information organized on paper notes into the system after the fact. At the same time, freeform pen input recognition is not yet mature or robust enough for deployment in mission-critical systems. Nevertheless, it is possible that freeform pen input can provide further insights into cognitive effort. In much the same way as we expected the quality of pen-gesture inputs and handwriting to be affected by high mental demand, we expected there to also be changes in low-level temporal features of freeform pen input to signify reduced resources available for cognitive processing [Ruiz 2011]. Indeed, our investigations reveal significant changes in stroke frequency to be correlated with low versus high cognitive load [Ruiz et al. 2011]. We also found that the discrepancy between stroke frequencies under low and high load is reduced with expertise. These results indicate that pen stroke frequency, which can be automated, could be used as an indicator of cognitive load, or conversely, of expertise level [Ruiz et al. 2011].

The analysis was carried out on a dataset of freeform pen input, generated as part of the same user study from which the pen-gesture input data came from, generating data in three levels of increasing load (sample inputs are shown in Figure 7). In contrast with symbolic pen-gesture inputs, however, freeform pen markings were used solely in

the digital notepad area and did not trigger any specific functionality. The role of the digital "notepad" was simply to emulate low-tech tools such as pen and paper.

Due to the high variability in content matter in each user's scratchpad, the analysis needed to be based on features sufficiently abstracted from the task content and semantics of the data itself. The features we investigated were based on the pen trajectories and the task time, chosen to ensure the content from all subjects could be judged equally. Using normalized stroke frequency measures (strokes per second), we found a main effect of cognitive load, where the frequency increased as cognitive load increased [Ruiz et al. 2011]. This signified that operators were writing much faster and relying on the digital notepad much more as cognitive load increased [Ruiz et al. 2011].

However, we also found that the discrepancy between stroke frequencies under low and high loads reduced with expertise, with both eventually converging. This suggests the possibility of using a convergence of stroke frequency in spite of varying task complexity to diagnose gains in expertise: increased expertise indicates an improvement in schema representations in working memory and the fact that learning has taken place [Ruiz et al. 2011]. Noting the fact that stroke frequency can easily be extracted in real time and unobtrusively using a tablet monitor or electronic pen, not only can this measure be applied to assess cognitive load levels, but also to detect when a user has acquired enough expertise on a given task or interface and hence allow them to progress to the next level of complexity [Ruiz et al. 2011]. An issue with using such methods to gauge improvements in expertise, however, is that efficacy of the index will be reduced with improved expertise, as subjects experience lower cognitive load; experts learn from the task and hence cognitive load variation occurs over a smaller range.

5. PROPERTIES OF BEHAVIORAL INDICES OF LOAD

Previous work on mental load measures, most notably by O'Donnell and Eggemeier [1986], Wickens and Hollands [2000], Kramer [1991], and Gopher and Braune [1984], has sought to describe them using a series of properties, to enable comparisons to be made such that the most appropriate measure is chosen for any situation. These include: sensitivity, diagnosticity, primary task intrusion, implementation requirements, operator acceptance, selectivity and bandwidth, and reliability.

In the context of our work, other properties have also proven useful in classifying behavioral indices, namely: the potential for implementing this measurement in real time; the provision and use of contextual information in interpreting the index or measure, dimensionality, and temporal scales.

Closely related is the issue of weighting methods for each of the individual modal index types and their subfeatures during fusion approaches. Weighting can be based on the task context, or can be based on the sensitivity or diagnostic power of the index, feature, or modality from which it is sourced [Ruiz 2011]. Confidence levels for the reliability of each index, as well as the combined multimodal index, can be provided on a task or user basis, depending on the quality of calibration and index combination types available [Ruiz 2011]. Other limitations can occur when combining data that is derived from inputs which vary sampling rates.

5.1. Real-Time Potential

One of the main goals of this work was to produce an automated method for cognitive load assessment that would allow the measure to have a high potential for being implemented and used in real time. Of course the definition of "real time" varies from one application to the next, for example, updates to the estimated cognitive load level at intervals of between 1ms to 10 minutes would be considered real time in the examples of Section 3.3. The basic requirement then is that the features used for assessment be extracted automatically, without the need for labeling or human

intervention. The features used to derive each of the modal indices presented here can be fully automated; the process from extraction to assessment can be done on-the-fly with very good results. Both types of indices (speech based and digital pen based) require some level of user calibration at the very beginning in order to improve the accuracy of the results. However, this process is quite simple in all instances and not at all prohibitive either in terms of time or effort. Individual measures within these modal index types may have differing levels of automation, for example, some may require a baseline or bootstrapping sequence at initialization.

5.2. Temporal Scales

Whether the features themselves offer discrete or continuous assessments will also affect how they are combined in a multimodal index. Information about how often the feature is updated or refreshed will need to be included as part of that single modality index information profile and serve as a reference as to which others it can be combined with. For example, in a constant speech-signal monitoring scenario, the acoustic speech index can be updated after every 2s of active speech or less. In contrast, the linguistic index based on the same raw input will have a longer update response lag, since the speech needs to be transcribed on-the-fly and a minimum amount of data needs to be accumulated before classification, for example, a complete sentence, phrase, or word. Hence, specific temporal update windows will need to be defined for specific modal index combinations. The granularity at which each type of index is refreshed or updated can be increased by implementing sliding window algorithms on the streaming signal, and initially weighted less heavily during the fusion stages. Many of the indices explored have displayed significant differences in average-based or rate-based features (e.g., MDIST, freeform pen), which means they have the potential to be indicative even with fewer samples than those collected here.

5.3. Dimensionality

Dimensionality refers to the number of features and individual measures included in a single index. For example, the MDIST measure is a single reading that combines the information from 12 separate information points derived from a trajectory. Particularly when using multidimensional indices, weighting of each subfeature can be a make or break factor, especially if the availability or quality of the sensor data on which these features is based cannot be guaranteed. Further, in early fusion implementations, the dimensionality of the features will certainly have an effect on how often they can be updated and combined with each other and contribute to the final multimodal index: features with high dimensionality and high update frequency can be revised much more often and potentially a higher level of confidence can be attached to such an index.

5.4. Contextual Information

Any multimodal index of load will need to be tightly coupled with the task context and workflow process. An understanding of the task flow is imperative because the indices presented here cannot be implemented as a universal solution applicable in all scenarios. For example, the linguistic indices are most useful in think-aloud data, or human-human communication between operators, but would be completely ineffective when applied on command-and-control speech input. Therefore, the contextual information will need to be closely derived from the work process, to select the indices most likely to be: (a) present, (b) reliably collected, and (c) the best match in each task scenario. In other cases, more than one modal indicator may be activated, for example, when an operator is speaking on the phone, the handwriting indices and a continuous speech signal index may be equally effective. Similarly, the contextual information can be considered on a per-user basis; user profiles can notify the system of individual

Table I. Properties Previously Defined in the Literature

Property	Definition
Sensitivity	Capability for discriminating significant variations along the workload continuum (Refer to Fig. 1 and Fig. 2.)
Diagnosticity	Capability for discriminating the specific computational process causing the load changes.
Primary Task Intrusion	Whether the task workflow is interrupted or not
Implementation Requirements	How difficult it is to implement within context, includes operator training or instrumentation required
Operator Acceptance	Willingness of operators to follow instructions.
Selectivity	Whether the measure is sensitive to mental workload only or also to physical changes.
Bandwidth and reliability	Bandwidth and reliability refer to the workload's estimate that has to be reliable both within and across tests.

Table II. Five More Pertinent Properties We Define

Property	Definition
Real-time potential	Is it possible to automatically extract these features from the input signal?
Temporal Scales	How often can this feature be updated?
Dimensionality	How many dimensions in this measure?
Contextual Information	Information regarding which indices can be combined with which others for maximum effect – and which should not be combined at all.
Domain independence	How tightly coupled is this measure to the domain it was observed in?

preferences (e.g., Operator A does not like to use pen input or handwriting; Operator B is quite reliant on pen for taking notes, but not interacting with the system or issuing commands).

5.5. Domain Independence

The advantage of the kinds of behavioral indices for cognitive load that are presented here is that they largely remain domain independent. For example, among the speech features, the acoustic indices appear to be relatively successful regardless of the domain (as shown by the variety of lab experiments conducted and the real-life case studies presented). The pen features, such as handwriting, for example, can also be applied in any domain, as long as the user is using a digital pen that is instrumented with the right sensors to capture the relevant features that comprise the index. Domain independence refers to how tightly coupled the features are to the domain in which they were observed; high domain independence means a loose coupling exists, whereas low domain independence means a tight coupling exists between dataset and domain.

5.6. Summary

Table I and Table II summarize the properties previously defined in the literature as well as the pertinent new ones we defined.

Table III and Table IV summarize the measures for the most pertinent previously defined properties and three of the new properties. Implementation requirements, operator acceptance, bandwidth and reliability, contextual information, and domain independence can only be assessed meaningfully with reference to a specific application domain, and hence are not included in this table. "Low-Normal-High" denotes possible values that can be taken by the feature, but not at the same time. Some features cannot detect low levels, for example. These tables are based on experiments detailed or referenced within this article, as well as general attributes of modalities. Some modalities may take one of multiple values, for example, pronouns are sensitive to

Table III. Speech-Based Measures and Indices

Feature	Sensitivity	Diagnosticity	Primary Task Intrusion	Selectivity	Real Time Automation	Temporal Scales	Dimensionality
Linguistic Features							
Pauses	Low, High	High	Low	Med	High – using voice activity detection	Significant pauses take around 0.3s. Sliding window every 5 seconds	∼3 individual features (total duration, frequency and avg. length)
Pronouns	Low, High	High	Low	High	Med – Dependent on recognizer	Word Level-Sliding window every sentence or phrase, using voice activity detection	Single
Complexity	Low, High	High	Low	High	High – Dependent on recognizer, use voice activity detection	Word, Sentence or Phrase level	∼3 measures of complexity
Category	Low, High	High	Low	High	High – Dependent on and word categories, Sliding window every task	Word level	2–10 significant categories
Valence	Low, High		Low	High	Med to High– Dependent on recognizer, Sliding window every task	Word level	2 significant categories
Acoustic Features							
Acoustic	Low, Nor-mal, High+	High	Low	Med	High	2–10 second windows yield excellent results	High >72 features

Low and High load, but being either singular or plural they cannot easily discriminate more than two classes. The tables are intended to illustrate that no single modality possesses the ideal set of attributes, and that a combination can instead provide the required accuracy.

6. MULTIMODAL INDICES OF LOAD

Given previous successes in finding features from pen and speech input that allow us to differentiate cognitive load levels for up to 3 levels of load, the next step is to apply a multimodal index of load that combines output from different sources. Correlations

Table IV. Other Indices and Measures

Feature	Sensitivity	Diagnosticity	Primary Task Intrusion	Selectivity	Real Time Automation	Temporal Scales	Dimensionality
Handwriting							
Velocity, Length, Pressure, Orientation, Altitude, Azimuth	Low, Normal, High+	Med to High	Low	High	High - Per stroke basis	High - Per stroke basis	~6 individual features per stroke
Frequency	Low, Nor-mal, High+	Med to High	Low	High	High – per stroke basis	Med to High– Dependent on segmenta-tion	~3 individual features per stroke
Symbolic							
MDIST	Low, High	Med to High	Low	High	High – Per symbol basis	High	~12 features per stroke
Custom, Geometric, Features	Low, High+	Med to High	Low	High	High – Per stroke basis, dependent on segmenta-tion scheme	High – Dependent on recognizer	Variable
GSR							
Mean	Low, Med, High +	High	Med – can use em-bed-ded sensors	Low	High – sampled every 100ms	High	Variable

between single-modality indices offer a way in which to introduce redundancy and robustness to a multimodal index of cognitive load. Dual-modality indices working together in a complementary fashion, such as speech-signal-based classification or degree of degeneration of pen input, are likely to align quite well, reinforcing each other. However, there are a number of aspects that need to be considered in the development of a multimodal index of load, for example, whether early or late fusion approaches are used. At an abstract level, multimodal indices can be derived in four ways [Ruiz 2011]:

(1) combining component features within each modality for, example, combining within pen-input features such as stroke frequency, MDIST, or altitude span;
(2) combining component features across modalities, for example, combining stroke frequency (from pen) with use of singular pronouns (linguistic);
(3) combining index results between modalities, for example, between pen-only assessment versus speech signal-only assessment;
(4) using a combination of any of the three preceding methods.

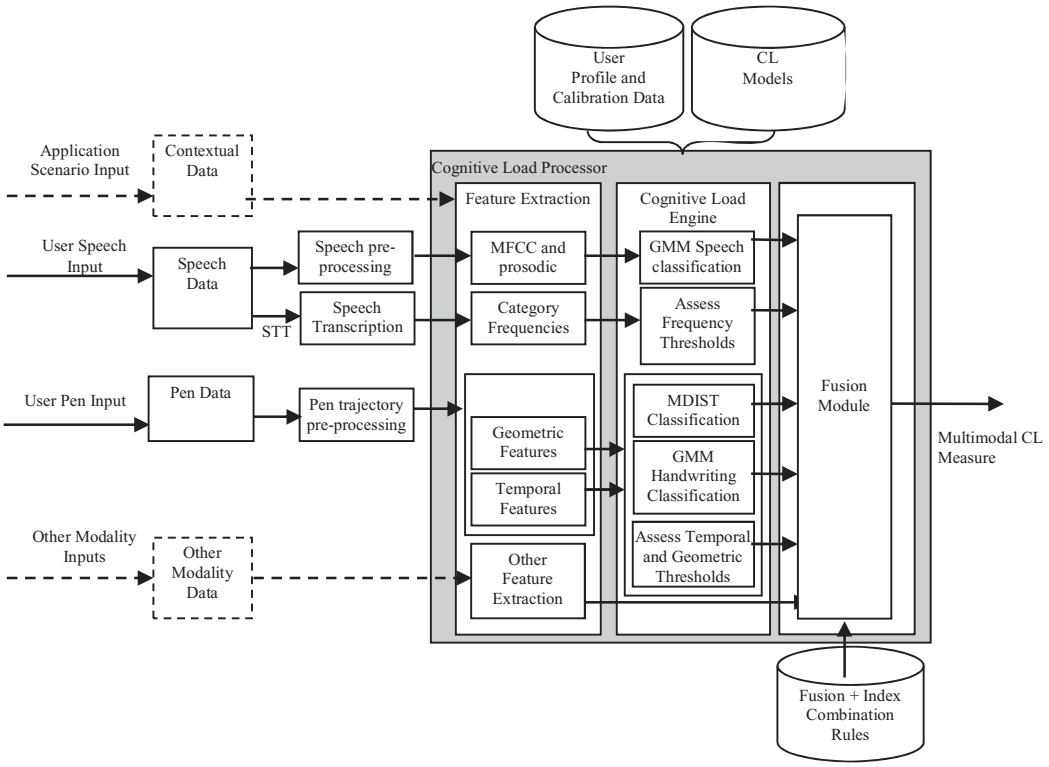


Fig. 8. High-level functional model of a multiple-modality CLM system.

6.1. An Abstract Model for Multimodal Assessment

Figure 8 depicts a high-level functional model of a proposed Cognitive Load Measurement (CLM) system. The abstract system model embodies four high-level processes: preprocessing and data cleaning, feature extraction, load assessment, and index fusion. The great advantage of multimodal behavioral indices of cognitive load is that they are derived from activity already undertaken as part of the task, and thus can be collected implicitly, or "passively" [Zander and Kothe 2011]. The raw modality input sources are first and foremost intended for purposes other than cognitive load measurement, specifically to do with the domain application.

For example, the data may be used for semantic interpretation or rendering (e.g., in the case of command-and-control speech or interactive pen gestures). The data may therefore need to be duplicated and diverted, with the original stream sent to the recognizers, and a secondary stream sent to the cognitive load measurement engine. In Figure 8, speech input data is first captured through a close talk microphone. This generates two kinds of data, speech signal data (e.g., a wav file) and text (through a speech to text engine). Likewise, pen input data is collected as trajectory tuples, including pressure, pen orientation, and other information transmitted directly from the device drivers, alongside system time-stamps.

Data preprocessing and data cleaning refers to any reformatting or restructuring of the input data, or removal of unnecessary information, for example, any outliers or segments that are too short for geometric and temporal analysis; words not recognized in the text, as well as words that are not used in the analysis. Input streams from other modalities will follow the same processes. Similarly, a number of other nonbehavioral

indices will also undergo preprocessing as needed; these include indices that may also be used in the process, such as galvanic skin response, or other body-based data, such as posture, movement, or temperature. Environmental and other external context information may also be provided to the CLM system for enhanced performance at this point.

The second stage involves streaming the individual modal inputs into their respective feature extraction components. The same data may be used for multiple feature extraction components, while other extraction components may not be activated, depending on domain-specific contextual information gathered from the active applications and workflow diagrams established a priori. This will allow the feature extraction engine to choose the most appropriate modules to activate for each incoming input stream. For example, if the incoming speech is sourced from a phone call or radio conversation, the feature extraction component will activate both MFCC and prosodic feature extraction as well as the linguistic category extraction components, since both can provide meaningful measures on this kind of data. On the other hand, if the incoming speech is sourced from command-and-control input, only MFCC and prosodic feature extraction will be activated, as the linguistic categories cannot provide any meaningful cognitive load measurement information on short, closed vocabulary, or single-word speech.

The third stage involves the decision-making aspect of the process, where thresholds are invoked and the appropriate models for each modality are selected from the database from which to carry out the classification. For example, for the speech-signal-based cognitive load measurement, different models are required for single word cognitive load classification versus continuous speech classification. Likewise, different MDIST models exist for each shape, and also for each user. Any calibration data that is needed for classification or for comparison purposes is also accessed at this point.

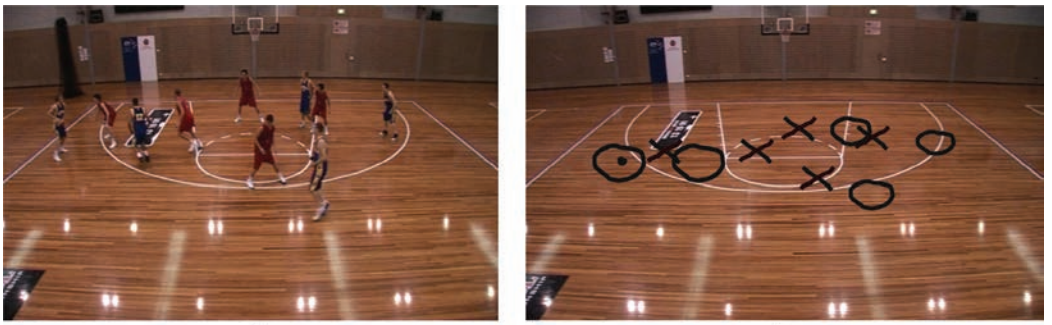
The final stage involves the fusion of indices resolved from the previous stage. The assessment results obtained from each modality can also convey confidence information to support the fusion process. The fusion engine accesses information regarding the modality load assessment combination rules in each specific context, for example, whether the time windows for the collected inputs are compatible; which indices are complementary with which others; and the appropriate weightings for each index, given the scenario and the user situation. Figure 8 shows how mid- and late- fusion may be achieved from a set of cognitive load assessments from each of the subfeatures. Mid-level fusion, for example, is achieved by combining multiple assessments that are based on the same input modality, for example, speech-based and linguistic assessments. Late fusion for a multimodal index can be likewise achieved by combining the results from all the features individually (regardless of input modality), or combining the input modality subgroup from the mid-level fusion results. The final output from the CLM engine can then be passed onto the output generation system in order to implement appropriate adaptation strategies.

We now present a user study illustrating the applicability of this model to multimodal data processing.

6.2. Basketball Skills Training

In order to illustrate how a multimodal cognitive load measurement system could work, we now present a lab-based study in which cognitive load and complexity were manipulated, and multiple behavioral modalities were recorded. The objective is to assess how well individual and combined modalities can reflect levels of cognitive load, and provide a concrete application for our multimodal cognitive load measurement model. While the task is different from the safety-critical, data-laden, and high-intensity applications discussed in the motivation for this article, it is a richly multimodal dataset that helps to provide an example application of the model proposed in Section 6.1.

Fig. 9. Physical setup of a user completing a task using a digital pen and with GRS attached.



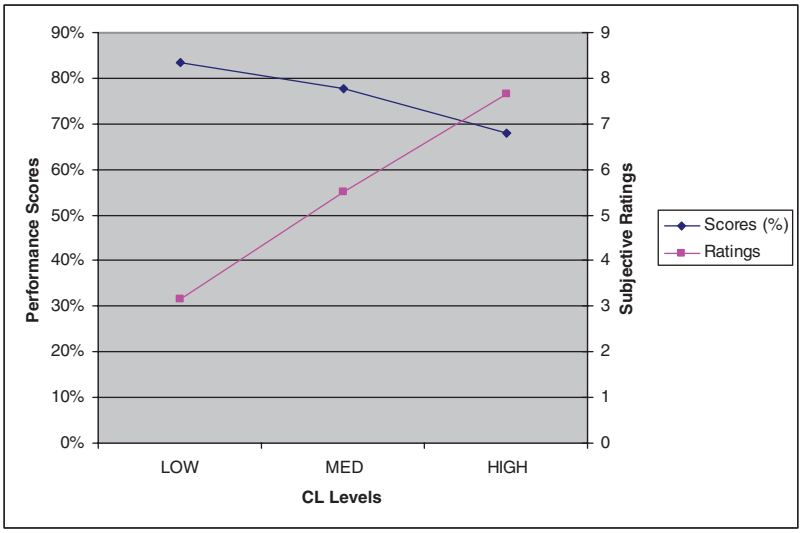
(a) (b)

Fig. 10. (a). Last frame of video clip before freeze. (b) Blank court image with player markings.

Elite athletes at the Australian Institute of Sport (AIS) are required to complete cognitive skills training using a targeted sports-specific software application called AISReact [Mackintosh 2010]. While aiming at ever-faster situation analysis and decision making through the construction of better mental schemas, it is desirable to precisely determine onsets of very high cognitive load in order to adapt the training rate to each individual athlete. In this experiment, we modified the software to accept pen-based interaction, and added the modalities of speech and eye activity. In addition, performance (accuracy) measures, physiological signals (GSR), and subjective ratings were also collected to establish a ground truth for cognitive load and task difficulty. The setup is shown in Figure 9.

Twelve male recreational basketball players, aged 19–36, each with more than 2 years' experience (average of 9.4 years) volunteered to complete the study. The task consisted of a 10s video basketball clip played on a tablet monitor, which was then frozen and replaced with a blank court schematic. The clips involved 10 players and the participants had to remember the locations and roles of some players in three task difficulty levels (remember 3 players for low level, 6 for medium, and all 10 for high). Each level consisted of 6 distinct clips. The clips were filmed from above and cover half the court, with all plays moving from the bottom of the screen towards the top, where the basketball hoop was located, as seen in Figure 10.

The participants used specific pen marks to identify the remembered player positions on the tablet monitor: attackers were denoted by crosses, defenders by circles, and



?

Fig. 11. Performance scores and subjective ratings [Ruiz et al. 2010].

the ball carrier by a circle with a dot in the middle, as illustrated in Figure 10(b). Participants were also instructed to think aloud through their answers, and these utterances were captured using a close talk microphone.

6.3. Subjective Ratings and Performance Results

Subjective ratings were collected using a Likert 9-point scale, where 1 was minimal effort and 9 was extreme effort. The task complexity levels induced extreme levels of load as reflected in the subjective ratings, increasing significantly as cognitive load increased, with mean averages of 3.2 (SD = 1.34), 5.5 (SD = 1.62), and 7.6 (SD = 1.23) for the low, medium, and high load tasks, respectively, in Figure 11. Due to the non-parametric dataset, this was verified using Friedman's $\chi 2$ test ($\chi 2(12,2) = 25.53$, $p < 0.001$), where low, med and, high were ranked 1.00, 2.04, and 2.96, respectively.

As expected, the participants' performance decreased significantly from low load to high load. Scores were given for each mark whose centroid was placed within a radius of 8% screen distance (in pixels) from the correct player position, as recommended by basketball experts at the Australian Institute of Sport, who also annotated the correct player positions on the schematic. The mean score for the low, med, and high load tasks 83.5% (SD = 11.63), 77.7% (SD = 12.26), and 68.1% (SD = 15.14). The decrease was verified through a repeated-measures ANOVA test ($F(2,22) = 4.84$, $p = .018$). Subsequent planned contrasts show a significant linear ($F(1,11) = 5.59$, $p = 0.04$, $r = 0.46$) to the 0.05 level, with a medium effect size. This is evident in Figure 11 also, where the performance decreases gradually between low and medium load levels and then more steeply from medium to high levels.

Overall, participants' performance decreased significantly, while their subjective ratings of load increased significantly, from low load to high load, validating that the responses elicited by these tasks are affected by extreme levels of cognitive load.

6.4. Individual Modalities

In this section, we analyze the capacity of individual modalities to classify load levels. In addition to speech and pen input, we present Galvanic Skin Response (GSR) although

Table V. Confusion Matrix of Three-Level Speech Classification

		Classified as		
		Low	Medium	High
Testing samples from	Low	100%	0%	0%
	Medium	40%	6%	54%
	High	15%	3%	82%

Table VI. Pen-Input Trajectory Features

Geometric feature	Description	Accuracy on test samples
Duration	Stroke duration, in milliseconds	**32.6%**
Length	Cumulative distance between sampled points along the trajectory	**40.7%**
Mean velocity	Mean velocity of the stroke trajectory, calculated point to point	**30.7%**
Mean acceleration	Mean acceleration of the stroke trajectory, calculated point to point	**37.0%**
Area	The area in pixels taken by the circle shape, enclosed by the trajectory	**36.3%**
First.Last	Distance between the first and last points of the trajectory	**33.3%**
Overlap ratio	The ratio of the overlapping distance between the first and last points of the trajectory to the total size of the shape.	**37.4%**

it is a physiological measure, because we compare the relative potential of these three modalities in the next section. In this analysis, a cognitive load estimate was made at the end of the task, that is, after around 1–5s.

Speech data was analyzed for all 12 subjects as described in Section 3 (and in Ruiz et al. [2010]), and the results use the average of the two evaluation folds, classifying into three predesigned load levels. As shown in Table V, low load achieved 100% accuracy, and high load 82% of testing samples. Interestingly, however, testing samples from the medium load level were mostly misclassified into either the low or high load, suggesting that no distinct pattern was captured. We suspect participants with subtly varied basketball skills and load capacity may have experienced slightly lower or higher loads in this level. The average accuracy for the 3 levels was 62.7%.

Unfortunately, due to corrupt collected signals from some of the GSR input sensor, and data losses caused by unexpected crashes in the software, only 9 subjects have complete data for the purpose of fusion, and hence this subset will be exclusively used for the remainder of this case study. For these 9 subjects only, the average speech classification accuracy drops slightly, to 61.8%.

Pen input was analyzed through a set of simple, objective features based on circling shapes drawn by the participants. Table VI summarizes the features and their individual accuracy at classifying load levels for the 9 subjects. The results range from 31% to 41% for the 3-level classification, that is, in some cases not always outperforming a random classification.

Although galvanic skin response is not a behavioral measure of cognitive load but a physiological one (i.e., it is not a voluntary reaction but a function of the autonomic nervous system), it was used as a ground-truth measurement for the study. Measured in micro-Siemens (μS), the signal was simply analyzed using an average measurement over the task period, yielding a classification accuracy of 64.4% over 3 load levels, across all 9 subjects, using a leave-one-out evaluation scheme.

Table VII. AdaBoost Weights for Speech and Pen Input Features

	Speech	Duration	Length	Velocity	Acceleration	Area	First-Last	Overlap
Weights	0.686	0.150	0.053	0.000	0.051	0.059	0.000	0.002

Table VIII. AdaBoost Weights for Speech, Pen Input Features and GSR

	Speech	Duration	Length	Velocity	Acceleration	Area	First-Last	Overlap	GSR
Weights	0.478	0.176	0.055	0.05	0.041	0.053	0.011	0.000	0.181

6.5. Multimodal Fusion

In this section we fuse the preceding features extracted from speech, pen input, and GSR using the AdaBoost boosting algorithm. Boosting [Freund 1995; Schapire 1990] is a general ensemble learning algorithm that creates an accurate strong classifier H by iteratively combining a number T of moderately inaccurate weak classifiers h_t. By definition, a strong classifier has high classification accuracy on the dataset, while a weak classifier's accuracy is just above that of a random guess. The final strong classifier can be defined as

$$H(x) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} \alpha_t h_t(x) \geq 0 \\ -1 & \text{otherwise} \end{cases},$$

where α_t is a weight coefficient. In simple cases, each weak classifier is attached to a feature, so the process of combining weak classifiers in Boosting is equivalent to a feature fusion process.

We used AdaBoost [Freund and Schapire 1997; Schapire et al. 1998], an adaptive version of boosting. Sample weights are all initially set equal, then refined iteratively during a training process. In order to select those features that are most discriminative of a given problem, in each iteration AdaBoost selects a new weak classifier h_t with the minimal weighted classification error with respect to the training sample weight distribution, which means the newly selected weak classifier can guarantee the more important samples (samples with higher weights) are classified correctly. Then the weights of incorrectly classified samples are increased, so in the next iteration, AdaBoost can focus on these incorrectly classified samples.

Table VII details the weights obtained for speech and the pen features. The average classification accuracy when fusing all these features is 64.1% on the testing samples, for the 3 load levels across all 9 subjects. It is noted that this represents a small improvement over the speech-only accuracy in Section 6.4. Cognitive load classification of freeform pen features is a challenging task (this is seen also in Table VI), however, the prospects for handwriting features are considerably more positive, as seen from the results discussed in Section 4.2.

Similarly, Table VIII details the weights obtained when fusing speech, pen features, and also GSR. The average classification accuracy is then 77.8% on the testing samples, for the 3 load levels across all 9 subjects.

Adding the GSR feature provides a significant improvement, supporting the benefits of feature fusion for workload detection. This case study is proposed as one implementation example of the model, however, the results indicate that other behavioral

features, yet to be explored, may be able to provide further multimodal cognitive load measurement accuracy.

7. DYNAMIC SYSTEM ADAPTATION BASED ON COGNITIVE LOAD INDICES

Interactive intelligent systems equipped with methods for unobtrusive, real-time detection of cognitive load and general cognitive load awareness should be able to adapt content delivery in more appropriate ways by sensing what the user is able to cognitively cope with at any given moment. Presentation and interaction strategies can be used to adapt the pace, volume, and format of the information conveyed to the user, depending on his or her individual cognitive load experience [Ruiz 2011]. For example, in the case of a real-life bushfire management control center scenario, the interaction system may be able to adapt many elements of the interface to decrease the cognitive load experienced by a user: from highlighting a critical computer screen or a specific information window, to sorting and prioritizing task checklists, to showing controlled reminders, to filtering email or SMS messages, to redirecting phone calls to the less cognitively loaded operators, the system has the power to subtly ease the user's cognitive demand [Khawaja et al. 2010, 2009].

Recent advances in the design of applications and user interfaces have promoted the awareness of the user context. It is crucial to establish a reliable indicator of cognitive load for each individual, by assessing which feature patterns are likely to occur at high or low levels of load on a case-by-case basis, given that there are large individual variations within a trend or pattern from one person to another. Many of the potential pen and speech indices summarized earlier need a relative baseline or standardization feature. Also, user preferences can also be used as the basis for response strategies for high cognitive load; some users may prefer to be overtly alerted to the system detecting their high load, while others may choose to let the system support them in a more autonomous way, for example, redirecting incoming calls to voicemail.

7.1. Performance Monitoring

The type of multimodal interaction environment we are targeting, featuring high-complexity, safety-critical tasks, could benefit from dynamic adaptation based on cognitive load assessment, as part of performance monitoring. Such complex work scenarios do not provide readily usable metrics for an operator's performance; instead, debriefings are used to assess the team's performance and address any undesirable outcomes. Cognitive load assessment can provide a real-time indicator of the load experienced by each operator: from this point, the system can be equipped to provide feedback to them, offer a warning, or suggest ways in which the system can "help". Other less technically-oriented solution strategies are possible, for example, where team leaders or managers manually redirect incoming incidents or incoming tasks to operators who may have more cognitive resources available to attend to them, while scaffolding others who are struggling to cope with demand.

7.2. Targeted Training

The use of cognitive load indices in intelligent environments, possibly in conjunction with performance and other measures, could provide an individually targeted learning experience. Interface and system learning environments are often aimed at a group level, while training programs seldom take into account individual differences in cognitive load during progression through increasingly complex material. Although some systems already exist that are able to cater for performance differences in training scenarios that can adapt slightly to accommodate these, it is generally acknowledged within the field of educational psychology that performance does not always accurately reflect the level of load. The latter, in fact, represents the subject's cognitive cost, for

example, cognitive resources spent, mental effort invested [Kalyuga 2007] to produce these results. By deploying cognitive assessment during training sessions, learners can benefit from a self-paced curriculum, supported by system recommendations as to when it may be appropriate to advance to the next module. This could potentially reduce training time and increase efficiency, with learners spending more time on material when necessary, and less otherwise.

8. CONCLUSION AND FUTURE WORK

The work presented here summarizes research aiming to measure cognitive load expended by human operators, especially using unobtrusive, real-time measurements. These are crucial for practical applications, where they can be used to optimize user interaction.

Previous research has tried to assess users' cognitive load using several methods including physiological, performance-based, and subjective measures. However, interactive intelligent systems lend themselves more to the collection of behavioral measures—in particular, modal inputs and communication—for cognitive load assessment. The goal is to measure a user's cognitive load implicitly and in real time so as to adapt systems to users affected by high cognitive load, easing the demand and avoiding stress, frustration, and errors. This work presented here has explored the viability of a number of behavioral modal data sources, especially from speech and pen input, to identify symptomatic cues of high cognitive load.

The feasibility of using user input and behavior patterns as indices of cognitive load is supported by experimental evidence. The benefits of this approach are that these measures can be collected implicitly, that is, by monitoring variations in specific modal features executed during day-to-day usage of interactive intelligent systems, thus overcoming problems of intrusiveness and increasing applicability in real-world environments. Moreover, using symptomatic cues of cognitive load derived from user behavior, such as acoustic speech signals, linguistic analysis of transcribed text, digital pen trajectories of handwriting and geometric shapes, can be supported by well-established theoretical frameworks, including O'Donnell and Eggemeier's workload measurement [O'Donnell and Eggemeier 1986], Sweller's Cognitive Load Theory [Sweller et al. 1998], and Baddeley's model of modal working memory [Baddeley 1992; Sweller et al. 1998], as well as McKinstry et al. [2008] and Rosenbaum's [2005] action dynamics findings.

Behavior-based cognitive load measurement also benefits from its very means of data collection. It doesn't require extra physical instrumentation of the user or environment, since the inputs it captures are part of the natural interaction required by the task. Moreover, the data is always available and current, so long as the user is interacting with the system or completing a task. Such real-time assessment of the user's cognitive load can then help achieve the ultimate goal of adapting information selection and presentation in a dynamic computer interface with reference to load. The development of standardized tasks to compare cognitive load measures would go a long way to achieving more definitive comparisons between indices.

Extensive investigations into a complete speech signal analysis for cognitive load measurement have culminated in the development of a fully functional automatic cognitive load assessment engine, able to produce a result in real time without manual intervention. Providing reliable speaker-independent measurement of cognitive load (85% accurate over 3 levels, without the need to create a model for each individual subject) for data collected using a close talk microphone incurring minimal cost. This would be a significant improvement in industrial environments where no cognitive load assessment technology currently exists. The changes in the user's voice that characterize high cognitive load occur at the acoustic and prosodic features of speech data, thus, the technology is able to make an accurate assessment regardless of the specific

words uttered, meaning of the message, or vocabulary used. Likewise, it is difficult for the user to consciously manipulate the assessment.

In regards to the linguistic analysis of the speech data, our studies show that the frequency of selected linguistic and grammatical features changed between low and high load tasks. We have successfully isolated a number of cognitive indices based on pause features, grammar features, language complexity features, and word category features such as emotive and agreement words. These indices are an ideal tool to complement current speech-signal-based results because they assess the content of the user's speech.

The results of our ongoing research also suggest that pen-input data produced under high cognitive load will also exhibit symptomatic characteristics, specifically in the structure, form, and manner of the trajectories generated in pen gesture, handwriting, and drawing. The findings demonstrate that the quality of interactive pen-gesture trajectories degrades as tasks become more complex; altitude, pressure, and orientation features show significant changes in handwriting produced in high load situations; and finally, the frequency of sketching, drawing, and other note-taking activities using a digital pen increases significantly in very difficult tasks compared to very simple tasks. Of these three pen-input measures, structural handwriting analysis has proven the most promising index of cognitive load. Strokes and interstrokes provide a comprehensive record of writing behavior, conveying rich insights into the cognitive load experienced by a writer. The overall classification accuracy showed that pen altitude, pen orientation, and pressure reflect cognitive load variations successfully, reaching 75% accuracy over three load levels.

These specific modal changes in modal and communicative behaviors when cognitive resources are scarce reflect a mental mechanism designed to extend working memory and reserve resources for problem solving strategies and processes. Despite significant evidence across a variety of domains and tasks (including some using psychology-based task designs that are well-known for inducing cognitive load) linking physical alterations to behavioral changes, the question of causality, where we can definitively link these changes to cognitive load, is still an open issue, and one we are actively investigating. Further work also remains to determine the shortest possible period of time from which cognitive load may be reliably estimated for each modality, since to date, both in the literature and in this article, decisions are usually made at the end of the task.

Finally, we proposed a high-level model of a system for assessment of cognitive load using a number of behavioral indices over two modalities: speech and pen. The real-time assessment of cognitive load provided by the system offers new potential for dynamic support and adaptive system behavior, promising to optimize the human-computer interaction throughput, and reduce the burden placed on limited human cognitive capabilities.

ACKNOWLEDGMENTS

We would like to thank our annotators, students and volunteers whose efforts are presented here. Special thanks to Guanzhong Li and Zhidong Li for their essential contributions.

REFERENCES

ALIBALI, M. W., KITA, S., AND YOUNG, A. J. 2000. Gesture and the process of speech production: We think, therefore we gesture. *Lang. Cogn. Process. 15*, 6, 593–613.

ARK, W. S., DRYER, D. C., AND LU, D. J. 1999. The emotion mouse. In *Proceedings of the 8th International Conference on Human-Computer Interaction (HCI'99): Ergonomics and User Interfaces.* Vol. 1, Lawrence Erlbaum Association, London.

BACKS, R. W. AND WALRATH, L. C. 1992. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Appl. Ergonom. 23,* 243–254.

BADDELEY, A. 2003. Working memory and language: An overview. *J. Comm. Disord. 36*, 189–208.

BADDELEY, A. D. 1992. Working memory. *Science 255*, 556–559.

BERTHOLD, A. AND JAMESON, A. 1999. Interpreting symptoms of cognitive load in speech input. In *Proceedings of the 7th International Conference on User Modeling (UM'99).*

BRENNER, M., SHIPP, T., DOHERTY, E., AND MORRISSEY, P. 1985. Voice measures of psychological stress: Laboratory and field data. In *Vocal Fold Physiology, Biomechanics, Acoustics, and Phonatory Control,* I. Titze and R. Scherer, Eds., 239–248.

BRUNKEN, R., PLASS, J. L., AND LEUTNER, D. 2003. Direct measurement of cognitive load in multimedia learning. *Educ. Psychol. 38*, 1, 53–61.

BYRNE, A. J., SELLEN, A. J., AND JONES, J. G. 1998. Errors on anaesthetic record charts as a measure of anaesthetic performance during simulated critical incidents. *Brit. J. Anaesth. 80*, 58–62.

CHALKER, S. AND WEINER, E. 1998. *The Oxford Dictionary of English Grammar*. Oxford University Press, New York.

CHANDLER, P. AND SWELLER, J. 1991. Cognitive load theory and the format of instruction. *Cogn. Instruct. 8*, 4, 293–332.

COWAN, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci. 24*, 1, 87–114.

DALE, R., ROCHE, J., SNYDER, K., AND MCCALL, R. 2008. Exploring action dynamics as an index of paired-associate learning. *PLoS ONE 3*, 3.

DECHERT, H. W. AND RAUPACH, M. 1980. *Towards a Cross-Linguistic Assessment of Speech Production*. Lang, Frankfurt.

DELIS, D. C., KRAMER, J. H., AND KAPLAN, E. 2001. *The Delis-Kaplan Executive Function System.* The Psychological Corporation.

DONNER, W. AND HANCOCK, J. T. 2011. Upset now? Emotion contagion distributed groups. In *Proceedings of the International Conference on Computer-Human Interaction (CHI'11).*

ESPOSITO, A., STEJSKAL, V., SMEKAL, Z., AND BOURBAKIS, N. 2007. The significance of empty speech pauses: Cognitive and algorithmic issues. In *Proceedings of the International Symposium on Brain, Vision and Artificial Intelligence (BVAI'07).* Lecture Notes in Computer Science, vol. 4729. Springer.

FERNANDEZ, R. AND PICARD, R. W. 2003. Modeling drivers' speech under stress. *Speech Comm. 40*, 1–2.

FLESCH, R. 1948. A new readability yardstick. *J. Appl. Psychol. 32*, 3, 221–233.

FRANKISH, G., HULL, R., AND MORGAN, P. 1995. Recognition accuracy and user acceptance of pen interfaces. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computer Systems (CHI'95).*

FREUND, Y. 1995. Boosting a weak learning algorithm by majority. *Inf. Comput. 121*, 2, 256–285.

FREUND, Y. AND SCHAPIRE, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci. 55*, 1, 119–139.

GALEN, G. P. AND VAN HUYGEVOORT, M. 2000. Error, stress and the role of neuron-motor noise in space oriented behaviour. *Biol. Psychol. 51*, 151–171.

GAWRON, V. J. 2000. *Human Performance Measures Handbook*. Lawrence Erlbaum Associates.

GOLDIN-MEADOW, S., NUSBAUM, H., KELLY, S., AND WAGNER, S. 2001. Explaining math: Gesturing lightens the load. *Psychol. Sci. 12*, 516–522.

GOPHER, D. AND BRAUNE, R. 1984. On the psychophysics of workload: Why bother with subjective measures? *Hum. Factors 26*, 519–532.

GUNNING, R. 1952. *The Technique of Clear Writing*. McGraw-Hill.

GUTL, C., PIVEC, M., TRUMMER, C., GARCABARRIOS, V. M., MDRISTSCHER, F., PRIPFL, J., AND UMGEHER, M. 2005. Adele (adaptive e-learning with eye-tracking): Theoretical background, system architecture and application scenarios. *Euro. J. Open Dist. E-Learn. 2.*

HANCOCK, J. T., GEE, K., CIACCIO, K., AND LIN, J. M. 2008. I'm sad, you're sad: Emotion contagion in cmc. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'08).*

HANSEN, J. H. L. 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Comm. 20*, 1–2, 151–173.

HOCKEY, G. R. J. 2003. Operator functional state as a framework for the assessment of performance degradation. In *NATO Advances Research Workshop on Operator Functional State and Impaired Performance in Complex Work Environments II.*

IKEHARA, C. S. AND CROSBY, M. E. 2005. Assessing cognitive load with physiological sensors. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05).*

IQBAL, S. T., ZHENG, X. S., AND BAILEY, B. P. 2004. Task-Evoked pupillary response to mental workload in human-computer interaction. In *Proceedings of the International Conference on Computer-Human Interaction (CHI'04)*.

JACOBS, S. C., FRIEDMAN, R., PARKER, J. O., TOFLER, G. H., JIMENEZ, A. H., MULLER, J. E., AND STONE, P. H. 1994. Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research. *Amer. Heart J. 128*, 1, 1170–1177.

JAMESON, A., KEIFER, J., MULLER, C., GROSSMANN-HUTTER, B., WITTIG, F., AND RUMMER, R. 2009. Assessment of a user's time pressure and cognitive load on the basis of features of speech. In *Resource-Adaptive Cognitive Processes*, M. W. Crocker and J. Siekmann, Eds., Springer, 171.

KALYUGA, S. 2007. Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educ. Psychol. Rev. 19*, 387–399.

KATZ, C., FRASER, E. B., AND WAGNER, T. L. 1998. Rotary-Wing crew communication patterns across workload levels. In *Proceedings of the RTO HFM Symposium on Current Aeromedical Issues in Rotary Wing Operations*.

KENNEDY, D. AND SCHOLEY, A. 2000. Glucose administration, heart rate and cognitive performance: Effects of increasing mental effort. *Psychopharm. 149*, 1, 63–71.

KERANEN, H., VAYRYNEN, E., PAAKKONEN, R., LEINO, T., KURONEN, P., TOIVANEN, J., AND SEPPANEN, T. 2004. Prosodic features of speech produced by military pilots during demanding tasks. In *Proceedings of the Fonetiikan Paivat Conference*.

KERR, B. 1973. Processing demands during mental operations: Memory and cognition. *J. Memor. Cogn. 1*, 401–412.

KETTEBEKOV, S. 2004. Exploiting prosodic structuring of coverbal gesticulation. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04)*.

KHAWAJA, M. A., CHEN, F., AND MARCUS, N. 2010. Using language complexity to measure cognitive load for adaptive interaction design. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'10)*.

KHAWAJA, M. A., CHEN, F., OWEN, C., AND HICKEY, G. 2009. Cognitive load measurement from user's linguistic speech features for adaptive interaction design. In *Proceedings of the International Conference on Human-Computer Interaction (INTERACT'09), Part I*. Lecture Notes in Computer Science, vol. 5726, Springer.

KIRSCHNER, F., PAAS, F., AND KIRSCHNER, P. A. 2009. Cognitive load approach to collaborative learning: United brains for complex tasks. *Educ. Psychol. Rev. 21*, 1.

KRAMER, A. F. 1991. Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance*, D. L. Damos, Ed., Taylor and Francis, London, 279–328.

LAJOIE, S. P. 2000. *Computers as Cognitive Tools: No More Walls*. Lawrence Erlbaum, Hillsdale, NJ.

LE, P., AMBIKAIRAJAH, E., EPPS, J., VIDHYASAHARAN, S., AND CHOI, E. 2011. Investigation of spectral centroid features for cognitive load classification. *Speech Comm. 53*, 4, 540–551.

LE, P., EPPS, J., AMIKAIRAJAH, E., AND SETHU, V. 2010a. Robust speech-based cognitive load classification using a multi-band approach. In *Proceedings of the APSIPA Annual Summit and Conference (APSIPA'10)*.

LE, P., EPPS, J., CHOI, E., AND AMBIKAIRAJAH, E. 2010b. A study of voice source and vocal tract filter based features in cognitive load classification. In *Proceedings of the International Conference on Pattern Recognition (ICPR'10)*.

LENNON, C. AND BURDICK, H. 2004. The lexile framework as an approach for reading measurement and success. http://www.Lexile.com.

LIPP, O. V. AND NEUMANN, D. L. 2004. Attentional blink reflex modulation in a continuous performance task is a modality specific. *Psychophysiol. 41*, 3, 417–425.

LIU, J., WONG, C. K., AND HUI, K. K. 2003. An adaptive user interface based on personalised learning. *IEEE Intell. Syst. 18*, 2, 52–57.

LIVELY, E., PISONI, D. B., SUMMERS, W. V., AND BERNACKI, R. 1993. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *J. Acout. Soc. Amer. 93*, 2962–2973.

MACKINTOSH, C. 2010. *AIS React Software v.6.6*. Australian Institute of Sport.

MARCUS, N., COOPER, M., AND SWELLER, J. 1996. Understand instructions. *Educ. Psychol. 88*, 1, 49–63.

MARSHALL, S. P., PLEYDELL-PEARCE, C. W., AND DICKSON, B. T. 2003. Integrating psychological measures of cognitive workload and eye movements to detect strategy shifts. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. Vol. 5, IEEE Computer Society, Los Alamitos, CA.

MCKINSTRY, C., DALE, R., AND SPIVEY, M. J. 2008. Action dynamics reveal parallel competition in decision making. *Psychol. Sci. 19*, 1, 22–24.

McLaughlin, H. G. 1969. SMOG grading: A new readability formula. *J. Read. 12*, 8, 639–646.

Mousavi, S. Y., Low, R., and Sweller, J. 1995. Measurement and analysis methods of heart rate and respiration for use in applied environments. *J. Educ. Psychol. 87*, 2, 319–334.

Muller, C., Grossmann-Hutter, B., Jameson, A., Rummer, R., and Wittig, F. 2001. Recognising time pressure and cognitive load on the basis of speech: An experimental study. In *Proceedings of the 18th International Conference on User Modeling (UM'01)*.

Nickel, P. and Nachreiner, F. 2000. Psychometric properties of the 0.1Hz component of hrv as an indicator of mental strain. In *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society (IEA'00/HFES'00)*.

O'Donnell, R. D. and Eggemeier, F. T. 1986. Workload assessment methodology. In *Handbook of Perception and Human Performance*, vol. 2, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, 1–49.

Oviatt, S. 1997. Multimodal interactive maps: Designing for human performance. *Hum.-Comput. Interact. 12*, 93–129.

Oviatt, S. 2006. Human-Centered design meets cognitive load theory: Designing interfaces that help people think. In *Proceedings of the ACM Conference on Multimedia*.

Oviatt, S. 2009. Designing interfaces that stimulate ideational superfluency. In *Proceedings of the Conference on Research Foundations for Understanding Book and Reading in the Digital Age: Implementing New Knowledge Environments*.

Oviatt, S., Coulston, R., and Lunsford, R. 2004. When do we interact multimodally?: Cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04)*.

Paas, F., Ayers, P., and Pachman, M. 2008. Assessment of cognitive load in multimedia learning: Theory, methods and applications. In *Recent Innovations in Educational Technology that Facilitate Student Learning*, D. H. Robinson and G. Schraw, Eds., 11–36.

Paas, F., Tuovinen, J. E., Tabbers, H., and Gerven, P. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol. 38*, 1, 63–71.

Paliwal, K. K. 1998. Spectral subband centroid features for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *J. Cogn. Engin. Decis. Making 2*, 2, 140–160.

Picard, R. 1997. *Affective Computing*. MIT Press.

Reck, R. P. and Reck, R. A. 2007. Generating and rendering readability scores for project Gutenberg tests. In *Proceedings of the Corpus Linguistics Conference*.

Reynolds, D. A. and Rose, R. C. 1992. An integrated speech-background model for robust speaker identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'92)*.

Rosenbaum, D. A. 2005. The Cinderella of psychology: The neglect of motor control in the science of mental life and behavior. *Amer. Psychol. 60*, 308–317.

Rubine, D. 1991. Specifying gestures by example. In *Proceedings of the 18th Annual ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'91)*. ACM Press, New York.

Ruffell Smith, H. P. 1979. A simulator study of the interaction of pilot workload with errors, vigilance, and decisions. NASA tech. memo, Moffett Field, CA: NASA Ames Research Center.

Ruiz, N. 2011. Cognitive load measurement in multimodal interfaces. Ph.D. dissertation, University of New South Wales, Sydney, Australia.

Ruiz, N., Liu, G., Yin, B., Farrow, D., and Chen, F. 2010. Teaching athletes cognitive skills: Detecting load in speech input. In *Proceedings of the 24th BCS Conference on Human-Computer Interaction (HCI'10)*.

Ruiz, N., Taib, R., and Chen, F. 2006. Examining the redundancy of multimodal input. In *Proceedings of the Annual Conference of the Australian Computer-Human Interaction Special Interest Group (OzCHI'06)*.

Ruiz, N., Taib, R., and Chen, F. 2011. Freeform pen input as evidence of cognitive load and expertise. In *Proceedings of the International Conference on Multimodal Interfaces*.

Ruiz, N., Taib, R., Shi, Y., Choi, E., and Chen, F. 2007. Using pen input features as indices of cognitive load. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI'07)*.

Schapire, R. E. 1990. The strength of weak learnability. *Mach. Learn. 5*, 2, 197–227.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist. 26,* 5, 1651–1686.

SCHILPEROORD, J. 2001. On the cognitive status of pauses in discourse production. In *Contemporary Tools and Techniques for Studying Writing*, T. Olive and C. M. Levy, Eds., Kluwer Academic Publishers, London.

SCHWARTZ, D. L. AND HEISER, J. 2006. Spatial representations and imagery in learning. In *Handbook of the Learning Sciences*, K. Sawywe, Ed., Lawrence Erlbaum.

SEXTON, J. B. AND HELMREICH, R. L. 2000. Analyzing cockpit communication: The links between language, performance, error and workload. *J. Hum. Perform. Extreme Environ. 5*, 1, 63–68.

SHI, Y., RUIZ, N., TAIB, R., CHOI, E., AND CHEN, F. 2007. Galvanic skin response (gsr) as an index of cognitive load. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*.

SPIVEY, M. J., GROSJEAN, M., AND KNOBLICH, G. 2005. Continuous attraction toward phonological competitors. *Proc. Nat. Acad. Sci. Unit. Stat. Amer. 102*, 29, 10393–10398.

STROOP, J. R. 1935. Studies of interference in serial verbal reactions. *J. Experiment. Psychol.*

SWELLER, J., MERRIENBOER, J., AND PAAS, F. 1998. Cognitive architecture and instructional design. *Educ. Psychol. Rev. 10*, 3, 251–296.

TOLKMITT, E. J. AND SCHERER, K. R. 1986. Effect of experimentally induced stress on vocal parameters. *J. Experiment. Psychol. 12*, 302–312.

URE, J. 1971. Lexical density and register differentiation. In *Applications of Linguistics,* G. Perren and T. Trim, Eds., Cambridge University Press, 443–452.

VANDERBERG, R. AND SWANSON, H. L. 2007. Which components of working memory are important in the writing process? *Read. Writ. Interdiscpl. J. 20*, 7.

WICKENS, C. D. AND HOLLANDS, J. G. 2000. *Engineering Psychology and Human Performance* 3^{rd} Ed. Pearson/Prentice-Hall, Upper Saddle River, NJ.

WILSON, G. F. AND RUSSELL, C. A. 2003. Real-Time assessment of mental workload using psychological measures and artificial neural network. *Hum. Factors 45*, 4, 635–643.

WOOD, C., TORKKOLA, K., AND KUNDALKAR, S. 2004. Using driver's speech to detect cognitive workload. In *Proceedings of the* 9^{th} *Conference on Speech and Computer (SPECOM'04)*. International Speech Communication Association Press.

YAP, T. 2011. Speech production under cognitive load: Effects and classification. Ph.D. thesis, University of New South Wales, Sydney, Australia.

YAP, T. F., AMBIKAIRAJAH, E., EPPS, J., AND CHOI, E. 2010a. Cognitive load classification using formant features. In *Proceedings of the IEEE International Conference on Information Sciences, Signal Processing and Their Applications (ISSPA'10)*.

YAP, T. F., EPPS, J., AMBIKAIRAJAH, E., AND CHOI, E. 2010b. An investigation of formant frequencies for cognitive load classification. In *Proceedings of the Annual Conference of the International Speech Communication Association (InterSpeech'10)*.

YAP, T. F., EPPS, J., AMBIKAIRAJAH, E., AND CHOI, E. 2011. Formant frequencies under cognitive load: Effects and classification. *EURASIP J. Adv. Signal Process.*

YAP, T. F., EPPS, J., CHOI, E., AND AMBIKAIRAJAH, E. 2010c. Glottal features for speech-based cognitive load classification. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'10)*.

YIN, B. AND CHEN, F. 2007. Towards automatic cognitive load measurement from speech analysis. In *Human-Computer Interaction, Interaction Design and Usability*, J. Jacko, Ed. Springer, 1011–1020.

YIN, B., CHEN, F., RUIZ, N., AND AMBIKAIRAJAH, E. 2008. Speech-Based cognitive load monitoring system. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*.

YIN, B., RUIZ, N., CHEN, F., AND KHAWAJA, M. A. 2007. Automatic cognitive load detection from speech features. In *Proceedings of the Australian Computer-Human Interaction Conference (OzCHI'07)*.

YEH, Y. Y. AND WICKENS, C. D. 1988. Dissociation of performance and subjective measures of workload. *Hum. Factors 30*, 111–120.

YU, K., EPPS, J., AND CHEN, F. 2011a. Cognitive load evaluation of handwriting using stroke-level features. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'11)*.

YU, K., EPPS, J., AND CHEN, F. 2011b. Cognitive load evaluation with pen orientation and pressure. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'11)*.

YU, K., EPPS, J., AND CHEN, F. 2011c. Cognitive load measurement with pen orientation and pressure. http://icmi11.forge.nicta.com.au/papers/MMCogEmS2011_Yu.pdf.

ZANDER, T. O. AND KOTHE, C. 2011. Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general. *J. Neur. Engin. 8*.

Received June 2011; revised June 2012; accepted July 2012

Human Factors: The Journal of the Human Factors and Ergonomics Society
http://hfs.sagepub.com/

Analysis of Collaborative Communication for Linguistic Cues of Cognitive Load

M. Asif Khawaja, Fang Chen and Nadine Marcus

Human Factors: The Journal of the Human Factors and Ergonomics Society 2012 54: 518 originally published online 20 January 2012
DOI: 10.1177/0018720811431258

The online version of this article can be found at:
http://hfs.sagepub.com/content/54/4/518

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Human Factors and Ergonomics Society

SPECIAL SECTION: Methods for the Analysis of Communication

Analysis of Collaborative Communication for Linguistic Cues of Cognitive Load

M. Asif Khawaja, University of New South Wales, Sydney, Australia, **Fang Chen**, National ICT Australia (NICTA) and University of New South Wales, Sydney, Australia, and **Nadine Marcus**, University of New South Wales, Sydney, Australia

Objective: Analyses of novel linguistic and grammatical features, extracted from transcribed speech of people working in a collaborative environment, were performed for cognitive load measurement.

Background: Prior studies have attempted to assess users' cognitive load with several measures, but most of them are intrusive and disrupt normal task flow. An effective measurement of people's cognitive load can help improve their performance by deploying appropriate output and support strategies accordingly.

Methods: The authors studied 33 members of bushfire management teams working collaboratively in computerized incident control rooms and involved in complex bushfire management tasks. The participants' communication was analyzed for some novel linguistic features as potential indices of cognitive load, which included sentence length, use of agreement and disagreement phrases, and use of personal pronouns, including both singular and plural pronoun types.

Results: Results showed users' different linguistic and grammatical patterns with various cognitive load levels. Specifically, with high load, people spoke more and used longer sentences, used more words that indicated disagreement with other team members, and exhibited increased use of plural personal pronouns and decreased use of singular pronouns.

Conclusion: The article provides encouraging evidence for the use of linguistic and grammatical analysis for measuring users' cognitive load and proposes some novel features as cognitive load indices.

Application: The proposed approach may be applied to many data-intense and safety-critical task scenarios, such as emergency management departments, for example, bushfire or traffic incident management centers; air traffic control rooms; and call centers, where speech is used as part of everyday tasks.

Keywords: cognitive load measurement, collaborative communication, language analysis, bushfire management, adaptive interaction and support

Address correspondence to M. Asif Khawaja, School of Computer Science and Engineering, University of New South Wales, NSW 2052, Australia; e-mail: asif.khawaja@nicta.com.au.

HUMAN FACTORS
Vol. 54, No. 4, August 2012, pp. 518-529
DOI:10.1177/0018720811431258

INTRODUCTION

Cognitive load (CL) refers to the amount of mental demand imposed on a person by a particular task and is associated with the limited capacity of the person's working memory and the ability to process novel information (Chandler & Sweller, 1991; Sweller, 1988). It is derived from the semantic or representational complexity of the task. However, the same task can affect different users in different ways and can induce levels of perceived cognitive load that vary from one user to another. This variation is attributable to many reasons, including level of domain expertise, age, and mental or physical impediments.

In complex and time-critical situations, users of an interaction system, especially those working collaboratively, can experience high cognitive demands, which can interfere with their ability to complete a task at an optimum performance level. These cognitive demands are caused either by the complexity of the task being carried out or by the complex design of an interaction system, as in multimodal or multimedia interfaces, which may contain inappropriate amounts of content delivered to users simultaneously (Mayer, 2001). For example, high-intensity control room work environments, such as for air traffic control, require operators to manage many such interfaces, switching from one application to another, often on multiple screens and in time-critical scenarios. Operators will frequently use radios or mobile phones, make and answer calls, and speak to their colocated colleagues while completing their tasks. This complexity can result in extremely high cognitive load and hinder the users' ability to perform their task.

An understanding of the users' current cognitive load will enable researchers to implement strategies to adjust the interaction system's response, presentation, and flow of interaction material and provide users with appropriate support as per their cognitive burden, helping

them complete tasks more effectively. Moreover, for complex collaborative tasks for which many users communicate with each other to solve task-related problems, understanding cognitive demands can be particularly helpful.

Background

Measuring a user's cognitive load robustly and in real time is not a trivial task. Many researchers have attempted to assess users' cognitive load using several methods, including physiological, behavioral, performance, and self-reporting subjective measures.

Historically, the most consistent results for cognitive load measurement have been achieved through self-reporting subjective measures (Brunken, Plass, & Leutner, 2003; Paas, Merrinboer, & Adam, 1994). These measures require users to introspect on their perceived level of cognitive load induced by various tasks by answering a set of assessment questions immediately after the tasks.

Performance-based approaches includes two techniques: primary task measurement, which is based on user's performance of the task being completed, and secondary or dual-task methodology, based on the performance of a second task that is performed concurrently with the primary task. Primary task measures include task completion times, speed or correctness, and critical errors (Gawron, 2000; Paas, Ayers, & Pachman, 2008). The dual-task approach has also been incorporated by several studies (Leyman, Mirka, Kaber, & Sommerich, 2004; Marcus, Cooper, & Sweller, 1996; Sweller, Merrienboer, & Paas, 1998; Wada, Iwata, & Tano, 2001) and can effectively be used to measure the degree to which the primary task requires working memory resources (Kerr, 1973).

The physiological approach of cognitive load measurement is based on the assumption that any changes in the human cognitive functioning are reflected in the human physiology (Kramer, 1991). The measures that have been used to show some relationship between participants' mental workload and their physiological behavior include heart rate and its variability (Mousavi, Low, & Sweller, 1995; Nickel & Nachreiner, 2000), brain activity (changes in oxygenation and blood volume, electrocardiography, electroencephalography; Brunken et al., 2003; Wilson & Russell, 2003), skin conductance (Jacobs et al., 1994; Shi, Ruiz, Taib, Choi, & Chen, 2007), and eye activity (blink rate, eye movement, pupillary dilation; Backs & Walrath, 1992; Lipp & Neumann, 2004; Marshall, Pleydell-Pearce, & Dickson, 2003).

However, most of those approaches can be physically or psychologically intrusive and may not allow implicit measurement. Moreover, they can sometimes be used only post hoc and may disrupt normal task flow. Although they may be useful approaches in research situations, they are often unsuitable for deployment in real-life applications.

Behavioral measures, in contrast, can provide an objective, nonintrusive, and implicit analysis of users' cognitive load, as they are based on data collected from the users while they complete the task, without their realizing that their behavioral data are being recorded. The user cannot manipulate the data (as in the case of subjective ratings) and can perform the task naturally without any interference. Behavioral features of cognitive load include eye blinking and movement (Gütl et al., 2005), mouse clicking and keyboard key presses (Ark, Dryer, & Lu, 1999; Liu, Wong, & Hui, 2003), and digital-pen gestures and usage patterns (Oviatt, 2006).

Most behavioral features of cognitive load have been extracted from users' speech signal data (Berthold & Jameson, 1999; Jameson et al., 2009; Keränen et al., 2004; Yap, Ambikairajah, Epps, & Choi, 2010; Yin, Chen, Ruiz, & Ambikairajah, 2008). Examples of such features include pitch, prosody, speech energy, and fundamental speech frequency. Some studies have reported an increase in the participants' rate of speech as well as speech energy, amplitude, and variability in high load conditions (Brenner, Shipp, Doherty, & Morrissey, 1985; Lively et al., 1993). Others have found peak intonation (Kettebekov, 2004) and pitch range patterns (Lively et al. 1993; Wood, Torkkola, & Kundalkar, 2004) to be related to high cognitive load. Pitch variability has also been shown to potentially correlate to cognitive load (Brenner et al., 1985; Wood et al., 2004).

Apart from speech signal features, linguistic and grammatical features may also be extracted from users' spoken language and analyzed

for patterns indicating high cognitive load. These features may include speech pauses, self-corrections, repetitions, response latency, and language usage, for example, use of different word categories and parts of speech, such as nouns and pronouns, and grammatical structures. Such features may be collected from users' spoken or written language and are highly unobtrusive, as the data can be collected without interrupting task flow. Some researchers have looked at linguistic features as indices of high cognitive load, including pauses (Berthold & Jameson, 1999; Khawaja, Ruiz, & Chen, 2008), word frequency, and use of first-person plurals (Sexton & Helmreich, 2000). Various other studies have also used linguistic features for purposes other than cognitive load measurement (Kramer, Oh, & Fussell, 2006; Rhee & Kim, 2001; Stirman & Pennebaker, 2001).

In this article, we study some novel linguistic and grammatical features of cognitive load and analyze various aspects of language use, including word selection, parts of speech, and grammar. These linguistic features can be used as cognitive load indices in domains in which speech or conversational transcripts are used as the main forms of input. They can also be fused with other speech, behavioral, and performance indices proposed by others to enhance the overall performance of a state-of-the-art multimodal cognitive load measurement system.

Aims of the Study

Australia is one of the most bushfire-prone regions in the world, and there are thousands of fires that need to be managed annually (Owen, Douglas, & Hickey, 2008). As the impact of climate change results in more extreme weather events (Flannigan & Wagner, 1991; Fried, Torn, & Mills, 2004; Hughes, 2003), fire and emergency service work is becoming increasingly important and needs to be well managed to save the communities from their effects. We present a study involving a real-life bushfire emergency management task carried out in the field by experienced emergency service operators working as a team in an emergency control room. The study was performed with the objective of analyzing bushfire management operators' natural speech for linguistic indices of cognitive load while they collaboratively perform bushfire management tasks of different complexities.

The overall aim of the study was to understand how people's spoken and linguistic behavior changes when working in a team collaboratively and when completing complex and high-cognitive-load tasks compared with low-load tasks. Capturing changes in communicative behavior will help us ascertain whether any members of the team experience high load. This knowledge, in turn, will enable any technology or tool that supports teamwork to adapt to those situations more intuitively so as to support team members' work processes.

Hypotheses

The rationale for our hypotheses is based on human cognitive models and working memory structure and its limitations (Atkinson & Shiffrin, 1968; Baddeley, 1992, 2000, 2003; Kintsch, Patel, & Ericsson, 1999; Sweller et al., 1998), which affect human language production processes and cause different linguistic patterns in various cognitive load conditions.

We hypothesize that while working collaboratively on high-cognitive-load tasks, participants will speak more with each other to manage the high task complexity. Although a few studies have shown a decrease in spoken communication with an increase in workload (e.g., Kleinman & Serfaty, 1989), we expect that focusing on more cognitive tasks that involve active thinking processes will result in more communication or "thinking aloud," especially when working in a collaborative team setting, resulting in an increased word count. This hypothesis has been confirmed by many other studies showing that in high-mental-load conditions, as things become more complex, team members communicate more and provide more information and/or explanations to each other as a strategy to deal with increased task complexity (Foushee & Helmreich, 1988; Jensen, 1986; Katz, Fraser, & Wagner, 1998; Oser, Prince, Morgan, & Simpson, 1991).

We also expect greater use of agreement expressions, for example, "OK" or "Agree," with low-load tasks and more disagreement among team members as task complexity increases. Our intuition is that people feel more

confident and responsible in easy situations than in difficult ones, when they tend to feel more reluctant. Also according to Baddeley's (2000, 2003) working memory model, when people experience high cognitive load and their separate audio and visual working memory resources are overloaded with the task itself, it may interfere with their ability to consciously understand and agree with what is being said by other team members, and people may unconsciously become reluctant, disagree, and/or take some time to understand what is being said.

We also wanted to see how people's individualistic and collectivistic communicative behavior changes in different cognitive load conditions. It has been found that when working in a group and handling difficult tasks together, people prefer to share their efforts to solve the problem (Kirschner, Paas, & Kirschner, 2009). Accordingly, we expect people to use personal pronouns differently in different cognitive load situations. A personal pronoun is a pronoun that substitutes for proper or common nouns and can be categorized into first-person singular, such as *I* or *me*; first-person plural, such as *we* or *us*; second-person singular or plural, such as *you*; and third-person singular or plural, such as *he*, *she*, or *they*. For our objective, we wanted to compare people's usage preference for singular versus plural personal pronouns. We hypothesize that in tasks of a collaborative nature, there will be an interaction between level of cognitive load and use of singular and plural personal pronouns; that is, as the task complexity (and so the cognitive load) increases, the use of singular pronouns will decrease and the use of plural pronouns will increase.

Again, we expect people to work together more to share the cognitive load as task complexity increases (Kirschner et al., 2009), and the difference between their individual and their collaborative working behavior may be visible from their pronominal usage preferences. Use of a singular personal pronoun implies a person's preference to work on his or her own, whereas use of a plural personal pronoun implies working together in a group. Therefore, considering the linguistic aspects of the team cooperation, we expect that when team members work together more than individually, they will use more plural (team) personal pronouns, such as *we*, *us*, or *our*, than singular (individual) pronouns.

It should be noted that although the second-person pronoun *you* has both singular and plural properties, semantically it is not possible for any automated analysis tool to discriminate between the two usage types. Therefore, our main focus of the analyses will remain to study the behavior of first- and third-person pronouns and their singular versus plural usage. Nevertheless, the use of the second-person pronoun *you* will still be analyzed to see how its behavior changes with different load levels. However, it is not expected to be a useful cognitive load index.

METHOD

A bushfire management study was performed by experienced bushfire operators working collaboratively in a computerized incident control room. The study requires the operators to carry out highly complex bushfire management tasks.

Task Design

The task involved a bushfire management exercise carried out by at least four members of an incident management team (IMT) working in an incident control center. The team members carried out 10 tasks, each about 5 hr in duration, in four states of Australia, including New South Wales, Victoria, Tasmania, and Queensland. Each IMT involved in an exercise comprised the following:

- incident control, for the management of all activities necessary for the resolution of an incident;
- planning, for the collection, analysis, and dissemination of information and the development of plans for resolution of an incident;
- operations, for the tasking and application of resources to achieve resolution of an incident; and
- logistics, for the acquisition and provision of human and physical resources, facilities, services, and materials to support achievement of incident objectives.

In the exercises observed for this study, three main team members or operators were studied, including incident controller (IC) and the officers in charge of the operations and planning functions. Logistics officers were excluded

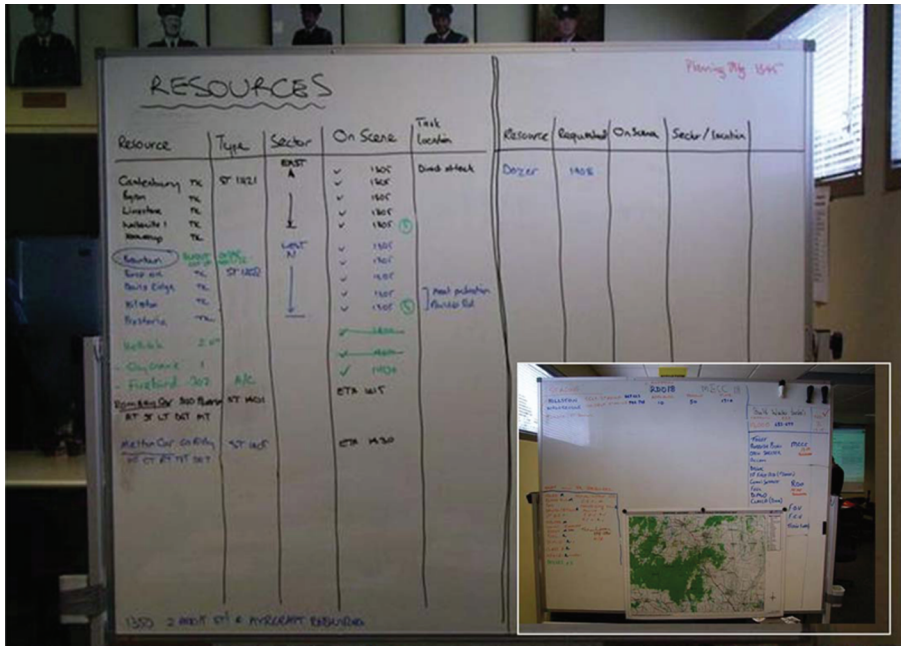


Figure 1. Use of charts and whiteboards by operators during an exercise at the training venue.
Source. Owen, Douglas, & Hickey (2008).

because of unavailability of sufficient data collection equipment.

All exercises were conducted as planned training exercises with the purpose of regular training for the bushfire management personnel. The exercises were conducted in genuine incident control centers. Thus, operators employed usual communication processes they would use in managing real bushfire incidents from the control centers. These processes included communicating information to each other and to other fieldworkers and volunteers via radios and telephones. Participants also used computers, often with multiple screens, for updated fire maps and task checklists and used paper-based reporting tools for updating the fire maps, charts, and boards with current fire status and resource information. Figure 1 shows a bushfire training venue and the use of charts and whiteboards by the participants during an exercise.

In the exercises, a wildfire is reported and an IMT is established. During the course of the exercise, the fire escalates and threatens local assets, such as a forest plantation and a town. The operators then perform the bushfire management activities from the incident control center. During the task, three levels of incident management activity occur randomly:

1. Low-level task demands, for example, little urgency and processes that are running smoothly.
2. Activity in which the incident escalates; for example, a change in strategy is needed as a result of deteriorating conditions, such as bad weather.
3. High level of task demand, for example, urgency and high resource coordination demands.

Each bushfire management exercise was facilitated and monitored by a bushfire management trainer in charge, who observed the operators' bushfire management activities and arbitrarily presented a subjective rating questionnaire to random operators to rate their experienced cognitive load.

Participants

A total of 33 male participants (11 teams of three operators) participated in the study. All participants had prior experience in firefighting, and the majority of them also had previous

experience in an IMT for bushfire management. All were trained in their role or function and were assumed to be competent for their IMT. All were native English speakers, so English speaking ability differences were assumed to be negligible. The exercises were aimed at personnel's being able to manage what develops to be a Level 3 incident, which is the highest level of bushfire incident complexity (Owen et al., 2008).

Data Collection and Coding Procedure

The three key roles (IC, operations, and planning officers) were video-recorded, and speech was also captured. We later transcribed the digital audio speech files collected using Transana (2010). Audio captured on the video recorders was also used to verify parts of the dialogue that were difficult to transcribe from the digital audio recorders. The transcribers were instructed not to include personal identification information. The speech transcriptions were then printed in hard copy and given to corresponding bushfire trainers in charge, who coded the transcriptions for cognitive load indication on the basis of their observations and available subjective ratings. The transcriptions were coded for four cognitive load levels according to the following framework developed by the bushfire training experts:

1. Low load (casual): Participants were involved in communication not related to their task, for example, conversation about personal life.
2. Medium load (routine): Participants were involved in nonchallenging routine bushfire management tasks.
3. High load (challenging): Participants were involved in challenging tasks, for example, handling unexpected events, producing information reports, and completing tasks within time constraints.
4. Very high load (very challenging): Same as high load; also participants were required to handle disturbances and breakdowns.

A partial sample of an operator transcription annotated with cognitive load is shown in Figure 2. Operators' coded transcripts were analyzed by at least three coders, who achieved an interrater reliability of 72%. The coders then discussed further the points of difference in an effort to reach the cognitive load coding framework as described, which resulted in 83% interrater reliability.

The electronic versions of transcriptions were then imported into NVivo (2010) and coded as per previous manual coding. Then the cognitive load–coded digital transcription file of each operator was processed semiautomatically to extract and separate the operator's transcription as per each cognitive load level and was saved. These saved files for each operator were then converted semiautomatically to a form usable by Linguistic Inquiry and Word Count (LIWC) automatic text analysis and extraction software (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007).

Analyses

We conducted detailed analysis of data collected from the study. During the preliminary analysis of the speech transcriptions for all four load levels, we observed that for the low and very high load levels, there were very few transcripts available, and they covered only 6 and 9 of the total 33 operators, respectively. To deal with this insufficiency of data, we merged the low-load data with medium-load data, and very-high-load data with high-load data, by averaging their corresponding participants' values. We referred to these resulting combined levels as *low load* and *high load*, respectively. Hence, all the statistical analyses and results that follow are based on the data for these two low- and high-load levels.

We used the LIWC tool to automatically extract the linguistic and grammatical features from the operators' speech transcriptions for both load levels. LIWC extracted the linguistic features as percentages of total words spoken by the operator to deal with operators' verbosity differences. The software counts the number of words for a specific linguistic or grammatical feature by matching the words from the transcription with its built-in dictionary of linguistic categories. The average dictionary coverage (the percentage of words captured by the

1	2	3	4	T3_Ops_1	5	6	7
				No, yeah, I just found it now.		2	
				[instructions			
				What do you want [change of Tactics]		2	
				(Offensive to defensive)			
				Yes, defensive		3	
				[, allocation of strike teams]			
				Yes. You don't have to tell him, Dave can tell him. Dave is doing a running between us now. That's right, he's already done it.		3	
				[sitrep from Firebird)			
				Yes, okay, Okay, well		3	
				(Response)			
				Yeah. Other than that what you've been told-it spotting all in here. yeah, we will lose all that in their		4	
				(Response)			
				I'll yeah, we'll lose it here			
				(Response)			
				It spotting into here now all right. Have we got time to back burn here. I'm not so sure now		3	

Figure 2. A sample bushfire transcription with cognitive load annotations. Note that the coding in column 6 represents the cognitive load level information entered by the bushfire in-charge coders.

dictionary) for the bushfire transcriptions was more than 86%. LIWC has been used by many other studies involving text and/or transcription analyses for purposes other than cognitive load measurement (Sexton & Helmreich, 2000; Stirman & Pennebaker, 2001).

RESULTS

Word Count

For all operators, the average word count for low-load task was 1,501.91 words, which is lower than that for the high-load task, 1,707.53 words, a difference of 13.69%. Although this trend is exactly as we hypothesized, the difference in the average word count between low-load and high-load levels was not statistically significant.

Words per Sentence

In line with the increased word count, we also expected longer sentences spoken by the operators resulting from more explanations happening with a high-load task. For all bushfire operators, average words per sentence for the low-load task was 9.21 words, which is found to be significantly fewer than that for the high-load task at 11.37 words, a significant difference of 23.44%, shown by a one-tailed t test ($t = –2.92$, $t_c = 1.69$, $df = 32$, $p = .003$). This finding suggests an increasing trend of words per sentence, confirming longer sentences and more communication with a high-load task than with a low-load task, as expected.

Agreement and Disagreement Words

We expected the bushfire operators to show more agreement among each other with low-load tasks than with high-cognitive-load tasks. To test this hypothesis for the study, the average number of agreement words used for the low-load task was found to be 3.48% of total spoken words, significantly higher than that for the high-load task at 2.01%, a difference of –42.07%, shown by a one-tailed t test ($t = 4.93$, $t_c = 1.69$, $df = 32$, $p = .0001$). Similarly, the average use of disagreement words for the low-load task is 1.39%, which is significantly lower than that for the high load task at 1.79% (significant difference of 29.19%), shown again by a one-tailed t test ($t = –2.47$, $t_c = 1.69$, $df = 32$, $p = .009$).

These results suggest that higher level of cognitive load is associated with decreased agreement and increased disagreement among people working together in a team to solve a problem.

Personal Pronouns

We conducted analysis of the use of personal pronouns to see whether there were any main effects and interaction between cognitive load levels and use of personal pronouns. We expected that bushfire operators would use more singular personal pronouns in a low-load task situation than with a high-load task. We merged first-person singular and third-person singular pronouns together and first-person plural and third-person plural pronouns together and then conducted the following analyses.

Main effects and interaction between cognitive load and personal pronouns. To check the main effects of the levels of cognitive load (low load vs. high load) and the pronoun types (singular vs. plural pronouns) and the interaction between the two, we performed a two-way repeated-measured ANOVA test.

We found a significant main effect of pronoun type (singular vs. plural pronouns), $F(1, 65) = 114.36$, $p < .001$. The operators' use of singular and plural pronouns, on average, was significantly different for low-load and high-load tasks. Specifically, operators used significantly more plural than singular personal pronouns in the high-load situation.

More importantly, the analysis showed that there was a significant interaction between the use of pronoun types (singular and plural) and cognitive load levels (low and high), $F(1, 65) = 18.33$, $p < .001$. This finding means that the effect of cognitive load level is different for singular and plural pronouns; that is, operators used more singular pronouns than plural in low-load situations and more plural pronouns than singular in high-load situations, as originally hypothesized. Figure 3 shows the significant main effect and the interaction between cognitive load levels and singular versus plural pronouns.

First-person singular and third-person singular pronouns. The aforementioned analyses showed significant interaction between singular and plural personal pronouns and levels of cognitive load. To further investigate the individual use of different singular pronouns, we conducted simple effects tests. For all operators, their

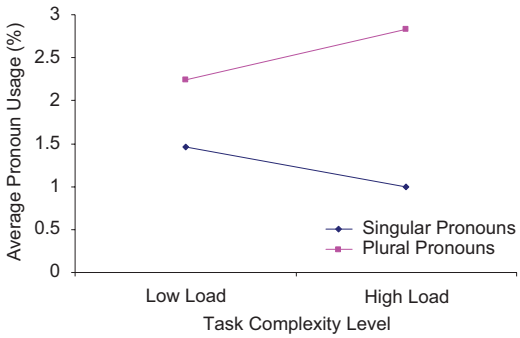


Figure 3. Main effects and interaction between pronoun types and cognitive load levels.

average use of first-person singular pronouns for the low-load task was found to be 2.21% of total words spoken, which is significantly higher than that for the high-load task at 1.6%, a difference of –28%, shown by a one-tailed t test (t = 1.96, t_c = 1.69, df = 32, p = .02). Similarly, average use of third-person singular pronouns for the low-load task is 0.699%, also significantly higher than that for the high-load task at 0.389% (difference of –44%), shown by a one-tailed t test (t = 2.27, t_c = 1.69, df = 32, p = .015). These results show a clear trend of overall decreased singular pronoun use linked with increased load.

First-person plural and third-person plural pronouns. For all operators, their average use of first-person plural pronouns for the low-load task is 3.20%, which is significantly lower than that for the high-load task at 3.97% (difference of 24%), shown by a one-tailed t test (t = –2.05, t_c = 1.69, df = 32, p = .02). Similarly, average use of third-person plural pronouns for the low-load task is 1.26%, significantly lower than that for the high-load task at 1.68%, a difference of 33%, shown again by a one-tailed t test (t = –2.28, t_c = 1.69, df = 32, p = .01). These results demonstrate participants' overall increased use of plural pronouns for high-load tasks, which shows the need to share the increased load among other team members.

Second-person pronoun you. As mentioned earlier, second-person pronoun *you* can be used for both singular and plural, so the results may be ambiguous. For all operators, we observed that the overall use of the second-person pronoun *you* decreased from 2.46% for the low-load task to 1.86% for high-load tasks, a significant difference of –24%. This result is similar to the behavior of other singular pronouns, but because of its inherent ambiguity in its singularity and/or plurality, it is difficult to conclude whether this difference was attributable to greater use of singular *you* pronouns for low-load tasks.

DISCUSSION

Analyses of some novel linguistic and grammatical features of cognitive load were carried out. The results confirm that while working collaboratively and performing high-cognitive-load tasks, people speak more with other team members to manage and share the high task complexity. The results show that participants, especially those working in a collaborative team environment, consistently use singular pronouns and plural pronouns differently in different task load situations. Specifically, they used significantly more singular pronouns for low-load tasks than for high-load tasks; that is, the lower the cognitive demand, the greater use of singular pronouns. In contrast, they used significantly more plural pronouns for high-load tasks than for low-load tasks; that is, the higher the cognitive load, the greater use of plural pronouns. These results support the notion that people actually collaborate and coordinate tasks more with each other during highly complex real-world tasks. These results are summarized in Table 1.

The results also suggest that in collaborative interaction situations, when dealing with low-cognitive-load tasks, team members are more confident about the task, prefer to perform tasks individually, and feel more comfortable accepting responsibilities and/or agreeing to the facts or instructions presented by other team members. In contrast, when dealing with complex and high-cognitive-load tasks, they do not agree easily and/or take individual responsibility; rather, they try to involve other team members to share the high and otherwise unmanageable cognitive load. This approach helps them to effectively solve problems during high-load tasks and improve the team's overall performance by working together and sharing the activities of the task when complexity increases

TABLE 1: Summary of Linguistic and Grammatical Features of Cognitive Load (N = 33)

Linguistic/Grammatical Features	Low-Load Task Average	High-Load Task Average	Difference*(%)
Word count[a]	1501.91	1707.53	13.69[c]
Words per sentence[a]	9.21	11.37	23.44
Agreement words[b]	3.48	2.01	−42.07
Disagreement words[b]	1.39	1.79	29.19
First-person singular pronouns[b]	2.21	1.60	−28
Third-person singular pronouns[b]	0.699	0.389	−44
First-person plural pronouns[b]	3.20	3.97	24
Third-person plural pronouns[b]	1.26	1.68	33

[a]Values in number of words.
[b]Values in percentage of total words spoken.
[c]Behavior as expected but not significant.
*p = .025 (one-tailed t test).

(Kirschner et al., 2009). Although these results apply across a variety of people, they may be specific to this combination of tasks in a collaborative bushfire management scenario. Further testing is suggested to confirm that results generalize to other types of applications, such as road or air traffic management tasks.

The linguistic assessment of cognitive load available through analysis of users' speech is attractive because it offers the potential to provide dynamic support and achieve adaptive system behavior, especially with the availability of appropriate technology for automatic speech recognition. If users experiencing high load can be identified by the system, they can be catered to with extra support, or perhaps through adaptation of the organizational or system behavior, to decrease their overall experienced cognitive load to more manageable levels. For example, in the bushfire management control center scenario, the system may be able to adapt many elements, such as highlighting a critical computer screen or a specific information window, sorting and prioritizing task checklists, showing controlled reminders, filtering e-mail or text messages, redirecting phone calls to the less cognitively loaded operators, and so on.

CONCLUSION AND FUTURE WORK

This study provides encouraging evidence and presents some novel linguistic and grammatical features extracted from natural speech as potential indices of users' experienced cognitive load. These features may be applied to many data-intense and safety-critical task scenarios, such as bushfire management or traffic incident management centers, air traffic control rooms, and call centers, where speech is used as part of day-to-day tasks on the phone or face-to-face. We expect that such promising features can complement other measures of cognitive load, such as physiological or performance features, and form part of a greater multimodal suite of measures acting together as robust indices of cognitive load.

We envisage a system that, after training, would be able to detect and calculate these and other similar speech and linguistic features automatically and to update these at regular intervals, such that an accurate indication of load is available at all times and can be used to update, modify, and adapt information presented to users in real time.

ACKNOWLEDGMENTS

We thank Christine Owen from Bushfire CRC, University of Tasmania, Australia, for providing the bushfire operators' data for this study.

KEY POINTS

- The manuscript proposes a linguistic analysis approach to human cognitive load measurement.
- Analysis of collaborative communication has been conducted for linguistic indices of cognitive load.

- Novel linguistic and grammatical features as measures of users' cognitive load have been identified.
- A new way of adapting system response and interaction and providing dynamic user support has been proposed.

REFERENCES

Ark, W. S., Dryer, D. C., & Lu, D. J. (1999, August). The emotion mouse. In H.J. Bullinger, & J. Ziegler (Eds.), *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction): Ergonomics and User Interfaces* (Vol. 1, pp. 818–823). London, UK: Lawrence Erlbaum.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195): New York, NY: Academic Press.

Backs, R. W., & Walrath, L. C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, *23*, 243–254.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Science*, *4*, 11, 417–423.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Journal of Neuroscience*, *4*, 829–839.

Baddeley, A. D. (1992). Working memory. *Science*, *255*, 556–559.

Berthold, A., & Jameson, A. (1999, June). *Interpreting symptoms of cognitive load in speech input.* Paper presented at the Seventh International Conference on User Modeling (UM99), Banff, Canada.

Brenner, M., Shipp, T., Doherty, E., & Morrissey, P. (1985). Voice measures of psychological stress: Laboratory and field data. In I. Titze & R. Scherer (Eds.), *Vocal fold physiology, biomechanics, acoustics, and phonatory control* (pp. 239–248). Denver, CO: Denver Center for the Performing Arts.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*, 53–61.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*, 293–332.

Flannigan, M. D., & Van Wagner, C. E. (1991). Climate change and wildfire in Canada. *Canadian Journal of Forest Research*, *21*, 66–72.

Foushee, H. C. & Helmreich, R. L. (1988). Group interaction and flight crew performance. In E. L. Weiner & D. C. Nagel (Eds.), *Human factors in aviation* (pp. 189-231). New York, NY: Academic Press.

Fried, J. S., Torn, M. S., & Mills, E. (2004). The impact of climate change on wildfire severity: A regional forecast for Northern California. *Journal of Climate Change*, *64*, 169–191.

Gawron, V. J. (2000). *Human performance measures handbook.* Mahwah, NJ: Lawrence Erlbaum.

Gütl, C., Pivec, M., Trummer, C., GarcaBarrios, V. M., Mdritscher, F., Pripfl, J., & Umgeher, M. (2005). Adele (adaptive e-learning with eye-tracking): Theoretical background, system architecture and application scenarios. *European Journal of Open, Distance and E-Learning*, *2*.

Hughes, L. (2003). Climate change and Australia: Trends, projections and impacts. *Journal of Austral Ecology*, *28*, 423–443.

Jacobs, S. C., Friedman, R., Parker, J. D., Tofler, G. H., Jimenez, A. H., Muller, J. E., . . . Stone, P. H. (1994). Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research. *American Heart Journal*, *128*, 1170–1177.

Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., & Rummer, R. (2009). Assessment of a user's time pressure and cognitive load on the basis of features of speech. In M. W. Crocker & J. Siekmann (Eds.), *Resource-adaptive cognitive processes* (p. 171). Berlin, Germany: Springer.

Jensen, R. S. (1986). *The effects of expressivity and flight task on cockpit communication and resource management* (Research Project 763247/714794, Grant No. NCC 2-206). Moffett Field, CA: National Aeronautics and Space Administration.

Katz, C., Fraser, E. B, & Wagner, T. L. (1998, October). *Rotary-wing crew communication patterns across workload levels.* Paper presented at the RTO HFM Symposium on Current Aeromedical Issues in Rotary Wing Operations, San Diego, CA.

Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T., Kuronen, P., Toivanen, J., & Seppänen, T. (2004). *Prosodic features of speech produced by military pilots during demandng tasks.* Paper presented at the Proceedings of Fonetiikan Päivät 2004, Oulu, Finland.

Kerr, B. (1973). Processing demands during mental operations. *Journal of Memory and Cognition*, *1*, 401–412.

Kettebekov, S. (2004, October). *Exploiting prosodic structuring of coverbal gesticulation*. Paper presented at the ICMI'04: 6th International Conference on Multimodal Interfaces, State College, PA.

Khawaja, M. A., Ruiz, N., & Chen, F. (2008, December). *Think before you talk: An empirical study of Relationship between speech pauses and cognitive load.* Paper presented at the Australasian Computer-Human Interaction Conference (OzCHI'08), Cairns, Australia.

Kintsch, W., Patel, V. L., & Ericsson, A. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*, 186–198.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). Cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review*, *21*, 31–42.

Kleinman, D. L., & Serfaty, D. (1989). *Team performance assessment in distributed decision making*. In R. Gibson, J. P. Kincaid, & B. Goldiez (Eds.), *Proceedings of the Interactive Networked Simulation for Training Conference* (pp. 22–27).

Kramer, A., Oh, L. M., & Fussell, S. R. (2006, April). *Using linguistic features to measure presence in computer-mediated communication*. Paper presented at the CHI2006, Quebec, Canada.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 279–328). London, UK: Taylor and Francis.

Leyman, E., Mirka, G., Kaber, D., & Sommerich, C. (2004). Cervicobrachial muscle response to cognitive load in a dual task scenario. *Ergonomics*, *47*, 625–645.

Lipp, O. V., & Neumann, D. L. (2004). Attentional blink reflex modulation in a continuous performance task is modality specific. *Psychophysiology*, *41*, 417–425.

Liu, J., Wong, C. K., & Hui, K. K. (2003). An adaptive user interface based on personalised learning. *IEEE Intelligent Systems*, *18*, 52–57.

Lively, E., Pisoni, D. B., Summers, W. V., & Bernacki, R. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America*, *93*, 2962–2973.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understand instructions. *Educational Psychology*, *88*, 49–63.

Marshall, S. P., Pleydell-Pearce, C. W., & Dickson, B. T. (2003). Integrating psychological measures of cognitive workload and eye movements to detect strategy shifts. In *Proceedings of 36th Hawaii International Conference on System Sciences (HICSS'03)* (Vol. 5, pages 6). Washington, DC: IEEE Computer Society.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Journal of Educational Psychology*, *87*, 319–334.

Nickel, P., & Nachreiner, F. (2000, August). *Psychometric properties of the 0.1Hz component of HRV as an indicator of mental strain*. Paper presented at the IEA 2000/HFES 2000: The XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society, San Diego, CA.

NVivo: Research Software for Analysis and Insight. (2010). Doncaster, Australia: QSR International. Retrieved from http://www.qsrinternational.com

Oser, R. L., Prince, C., Morgan, B. B., Jr., & Simpson, S. S. (1991). *An analysis of aircrew communication patterns and content* (Tech. Rep. 90-009). Orlando, FL: Naval Training Systems Center, Human Factors Division.

Oviatt, S. (2006, October). *Human-centered design meets cognitive load theory: Designing interfaces that help people think*. Paper presented at the MULTIMEDIA '06, New York, NY.

Owen, C., Douglas, J., & Hickey, G. (2008, May). *Information flow and teamwork in incident control centers*. Paper presented at the 5th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Washington, DC.

Paas, F., Ayers, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning: Theory, methods and applications. In D. H. Robinson & G. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 11–36). USA: IAP Inc.

Paas, F., Merrinboer, J. J. G. V., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*, 419–430.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Retrieved from http://www.liwc.net

Rhee, K. M., & Kim, E. (2001, August). *A statistical analysis of text for inferring authenticity*. Paper presented at the 53rd Session of International Statistical Institute, Seoul, Korea.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communication: The links between language, performance, error, and workload. *Journal of the Human Performance in Extreme Environments*, *5*, 63–68.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007, April). *Galvanic skin response (GSR) as an index of cognitive load*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI'07), San Jose, CA.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Journal of Psychosomatic Medicine*, *63*, 517–522.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Sweller, J., Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296.

Transana. (2010). Madison: University of Wisconsin–Madison Center for Education Research. Retrieved from http://www.transana.org

Wada, F., Iwata, M., & Tano, S. (2001). Information presentation based on estimation of human multimodal cognitive load. In *Proceedings of IFSA World Congress and 20th NAFIPS International Conference* (Vol. 5, pp. 2924–2929). Vancouver: IEEE.

Wilson, G. F., & Russell, C. A. (2003). Real-time assessment of mental workload using psychological measures and artificial neural network. *Human Factors*, *45*, 635–643.

Wood, C., Torkkola, K., & Kundalkar, S. (2004, September). *Using driver's speech to detect cognitive workload*. Paper presented at the 9th Conference on Speech and Computer (SPECOM'04), St. Petersburg, Russia.

Yap, T. F., Ambikairajah, E., Epps, J., & Choi, E. (2010, May). *Cognitive load classification using formant features*. Paper presented at the IEEE International Conference on Information Sciences, Signal Processing and Their Applications (ISSPA'10), Kuala Lumpur, Malaysia.

Yin, B., Chen, F., Ruiz, N., & Ambikairajah, E. (2008, March). *Speech-based cognitive load monitoring system*. Paper presented at the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08), Las Vegas, USA.

M. Asif Khawaja is a researcher at National ICT Australia, Sydney, as well as at the School of Computer Science and Engineering at the University of New South Wales (UNSW), Sydney, Australia. He completed his PhD in computer science from UNSW in 2010.

Fang Chen is a professor at UNSW and research group manager at National ICT Australia in Sydney. She completed her PhD in communications and electronic systems from Beijing Jiaotong University in 1994.

Nadine Marcus is a senior lecturer in human-computer interaction at the University of New South Wales (UNSW) in Sydney, Australia. She completed her PhD in educational psychology from UNSW in 1998.

Date received: January 14, 2011
Date accepted: October 17, 2011