Exploring Novel Methodology for Classifying Cognitive Workload

Seth Siriya[1(✉)], Martin Lochner[2], Andreas Duenser[2], and Ronnie Taib[3]

[1] University of Melbourne, Melbourne, Australia
ssiriya@student.unimelb.edu.au
[2] Data61, CSIRO, Hobart, Australia
[3] Data61, CSIRO, Eveleigh, Australia

Abstract. This paper describes our work in extracting useful cognitive load classification information from a relatively simple and non-invasive physiological measurement technique, with application in a range of Human Factors and Human-Computer Interaction contexts. We employ novel methodologies, including signal processing, machine learning and genetic algorithms, to classify Galvanic Skin Response/Electrodermal Activity (GSR/EDA) signals during performance of a customised game task (*UAV Defender*) in high- and low-workload conditions. Our results reveal that Support Vector Machine Linear was the most successful technique for classifying the level of cognitive load that an operator is undergoing during easy, medium, and difficult operation conditions. This methodology has the advantage of applicability in *critical task* situations, where other cognitive load measurement methodologies are problematic due to sampling delay (e.g. questionnaires), or difficulty of implementation (e.g. other psych-physiological measures). A proposed cognitive load classification pipeline for real-time implementation and its use in human factors contexts is discussed.

Keywords: Cognitive load · Galvanic Skin Response · Electrodermal Activity · Psycho-physiology · Analytics · Machine learning · Decision response task

1 Introduction

When operating complex machinery, dealing with sensitive control apparatus, or navigating vehicles of any size, it is increasingly evident that the level of cognitive load (CL) the operator is incurring has a direct impact on the operator's performance on the task, (e.g. [12]). Likewise, such multi-tasking during critical task operation has detrimental effects on both the primary task (e.g. driving) and secondary task performance [5,13].

Managing cognitive load promises to optimise the way information is processed and responded to by humans, addressing errors due to overload, which

P. Cortez et al. (Eds.): INTETAIN 2018, LNICST 273, pp. 105–114, 2019.
https://doi.org/10.1007/978-3-030-16447-8_11

was identified as a factor in the tragic crash of flight AF447 in 2009 [6], or conversely to underload, in typical surveillance scenarios where an operator must be able to detect minute anomalies in very long sequences of otherwise normal observations.

Where our current work has useful, and indeed extremely relevant scope given recent and expected technological advances, is in the domain of applied, mission-critical environments. In this context, we define a critical task as one that: (a) needs active and substantial attentional resources for its successful execution, and (b) could result (or will likely result) in catastrophic circumstances, i.e. injury, death, or damage to property, in the event of task failure. While the large body of existing laboratory evidence is useful in understanding CL, and predicting the effects of high and low workload tasks, there remains a gap in our ability to monitor this phenomenon in the actual operating environments where it is arguably most important. In addition to affording new ways of monitoring CL in real-time and measuring or evaluating operator performance and interaction with systems, psycho-physiological measures such as those described in this paper can contribute to the development of novel, real-time, direct and indirect human-computer interaction (HCI) techniques (e.g. [18,19]).

2 Related Work

Cognitive load corresponds to the mental effort expended carrying out a task, based on the premise that working memory capacity is a limited resource in the human cognitive system, yet is critical in coordinating memory, attention and perception [2].

By measuring the cognitive load experienced by a user, applications could adapt the amount and pace of content they provide to continuously optimise delivery, hence maximising the throughput of information between the human and computer. From the literature, we can identify four broad ways to measure cognitive load: subjective assessment methods, performance based techniques, behavioural measurements, and psycho-physiological measurements. We here focus on the latter.

Psycho-physiological measurements use changes in a person's physiological state to infer a change in mental state - in this case, cognitive workload. There are numerous examples of such measures, including direct measures of brain activity (EEG, FNIR; e.g. [1,7]), ocular activity (e.g., [19]), breath rate, heart rate speed and variability [23]. For a detailed review, please see [3].

In this paper, we focus on one of the oldest and most studied measures of human psycho-physiology: Galvanic Skin Response, also known as Electrodermal Activity. Whereas most of the above techniques require expensive apparatus, considerable time for set-up, and specific laboratory techniques for data collection, the GSR is of minimal complexity, requiring only two proximally-located electrodes in contact with the epidermis. The metric itself is simply the electrical conductance between these two points, as measured in microsiemens. Whereas historically the raw value of the GSR has been of primary interest, we present here some novel methodologies for usefully classifying these signals under different CL conditions.

Early work on GSR discriminated between tonic (baseline level, also known as skin conductance level) and phasic (fluctuations due to physiology activity, also known as skin conductance responses) GSR [11]. On the other hand, a study into driver stress simply smoothed the GSR signal with a digital elliptical filter cutting at 4 Hz, and then used slope characteristics (magnitude and durations) as features [8]. Another study specifically targeting CL measurement from GSR showed that a simple, unimodal metric such as *accumulated GSR* can be a reasonable indicator of CL [20]. Other research has shown that GSR can be a good indicator of the quality of human decisions. In this work the raw signal was first smoothed using a Hann window function, followed by z-standardisation before applying extrema-based and statistic-based features similar to the above studies [24]. Finally, some other studies have explored the applicability of feature extraction methods used for other signals to GSR. For example, EEG and EMG methods have been included for analysis in this investigation [9,15].

3 Methodology

45 Participants (ages 21–35) from a University research pool were involved in this study. They were paid 40USD for completing two 1-hour test sessions. Participants reported having normal or corrected-to-normal visual acuity.

Participants completed a computer-based task, UAV Defender, over two testing sessions. The task was developed at the Tasmanian Cognition Lab (University of Tasmania) and was carried out on standard current-model Windows desktop computers, and standard peripheral devices. The GSR signal was collected from the finger and thumb of the non-dominant hand, using commercially available Neulog GSR logger-sensors. GSR data was collected at 20 Hz. The ISO Decision Response Task [10] was implemented using the DRT kit available from RED Scientific (USA), using the haptic-buzzer stimulus setup, with the buzzer located at the left collar-bone, and a foot switch for reaction time (RT) responses.

The UAV Defender task requires participants to track multiple UAVs as they traverse a landscape, as viewed from a *birds-eye* viewpoint, i.e., from directly overhead. Similar in implementation to the Multiple Object Tracking task [17], we manipulated difficulty in three levels by requiring all participants to track either 3, 5 or 7 targets simultaneously. The level of difficulty was fully counterbalanced across trials. Participants were required to click on a UAV when a visible fuel level marker became low, as indicated by a colour bar on the UAV icon. Trials of UAV Defender lasted 2 min each, and participants completed 24 trials on two consecutive testing days, for a total of 48 trials (with random counterbalanced sequence of difficulty levels).

Data Cleaning. Of the initial 46 participants in this experiment, one participant was dropped completely due to an intermittent short in the GSR signal. For the remaining 45 participants, we implemented a data cleaning procedure to remove trials in which more than half of the trial data exceeded specific cut-off limits. High cut-off was 9.9 μS, and the low cut-off was 1 μS.

Metrics. While much of the past research into GSR and workload has generally studied the raw GSR signal (i.e. skin conductance), we have attempted here to develop a suite of metrics that are intended to increase our ability to differentiate high-workload GSR signals from low-workload GSR signals. These include both successful metrics from recently published work, and novel experimental metrics.

Standardized GSR and Accumulated GSR: Previously employed with success in [14], z-score standardization of the GSR signal has been shown to improve discriminability between workload levels. Accumulated GSR is the sum of raw signal values in a trial, which has been found as descriptive of CL [14].

Slope, or Gradient of the GSR Signal: Novel to our present analysis, this metric takes the gradient, or slope, of the GSR signal by taking the line of best fit over a rolling window of 40 samples (i.e. 2 s of data at 20 Hz). This methodology enables us to remove the overall drift component of the raw GSR signal, and smooth out high-frequency signal while retaining some local information, with a score of zero indicating no change in the GSR signal.

Zero Crossings: The rate of 'zero crossings', measures the rate at which the gradient changes from positive to negative. This metric enables us to assess the *speed* of the waveform without resorting to Fourier analysis such as the Fast Fourier Transform (FFT) or the Discrete Fourier Transform (DFT).

Negative Slope Percentage: Negative Slope Percentage (NSP) measures the proportion of time within a trial that the slope of the raw GSR signal is below zero, indicating a decline in skin conductance. This was an exploratory feature testing the hypothesis that decreased load results in decreasing skin conductance.

PSD Coefficients: Power Spectral Density (PSD) coefficients are descriptive of how important a frequency is for a signal. We used Welch's method to obtain the coefficients, since it allows control of the variance of the estimate (at the cost of frequency resolution) [22]. The spectrum for GSR is concentrated below 0.5 Hz [16], therefore we only analyse coefficients within this range. Then segpoints is defined as the number of PSD coefficients located between 0 and 0.5 Hz. Segpoints values that were investigated were 10, 20, 30, 40, 50 and 60. PSD coefficients were obtained from the Standardised GSR and slope.

Hjorth Parameters: We extracted the Hjorth parameters (which were developed for EEG analysis) from both the Standardised GSR and Slope of the GSR signal [9]. The parameters were extracted as features for both the Standardised GSR and the slope.

EMG-Based Metrics: Alongside the Hjorth parameters which were selected due to their success in EEG analysis, we also explored EMG-based features due to their usage of frequency-domain information. These were the first, second and third spectral moments, mean frequency, peak frequency and total power [15]. Again, features were extracted for both the Standardised GSR and slope.

Feature Exploration. To determine how descriptive these features are of cognitive load, we performed a one-way within-subject ANOVA test for each of the metrics on day one, with game difficulty level as the independent variable. We applied within-subject standardisation for the features to account for between subject differences. This is similar to feature calibration previously applied in CL classification research, but with z-score standardisation rather than divide-by-mean calibration due to the mean of some features being zero [14].

Classification Model. Following the analysis of features, we generated various classification models and evaluated their performance.

Model Selection, Training and Testing: The purpose of model selection is to find features (with within-subject standardisation) and parameters best suited to the problem of classifying CL. This was repeated on both days individually to get a unique set of features and model parameters corresponding to day one and day two. Model training was repeated for both days, such that a model was trained using parameters and features from day one, and repeating this separately for day two. Finally, models we trained on day one data are tested on day two data, and vice-versa. The purpose of testing on the opposite dataset is to observe how well models generalise to unseen data [4]. Model performance was measured using F1-Score.

Genetic Algorithm-based Feature Selection: The goal of feature selection during model selection is to find the best subset of features that are informative of CL levels. Since it would take a long time for an exhaustive search through all possible combinations of features, we applied genetic algorithm-based search to explore the feature space for each segpoints value.

Two classification algorithms were used in model selection: Naïve Bayes and K-Nearest Neighbours. Fitness scores were calculated via leave-one-subject-out cross validation [14]. We implemented a genetic algorithm in Python using the DEAP evolutionary computation framework, with most parameters replicated from previous work [21]. The sole exception was probability of mutation, which we increased to 0.01 due to the smaller feature space. Finally, we used three different learning algorithms for the model training/testing stages: K-Nearest Neighbours, Random Forest and Support Vector Machines (SVM) Linear.

4 Results

4.1 Feature Exploration

Generating PSD with multiple segpoints is computationally expensive and each PSD represents similar information (with slight differences due to variance-resolution trade-off). Therefore the best segpoints value should be determined. We approached this from the perspective of determining the segpoints value with the most number of significant features. Following ANOVA testing on all of the features, we counted the number of significant frequency-domain features for each segpoints, which is recorded in Fig. 1.

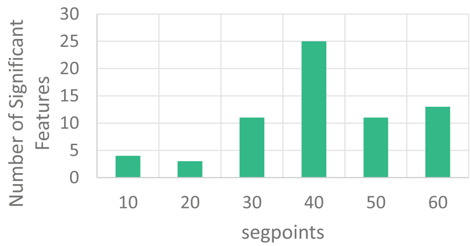
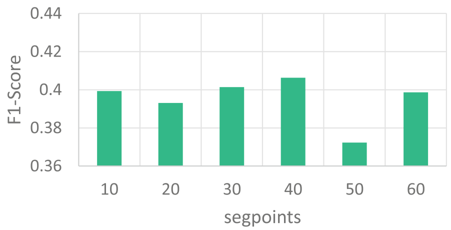


Fig. 1. Comparison of number of significant frequency domain features generated from PSD of varying segpoints

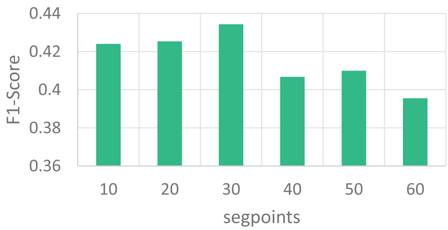
We also compared the individual features themselves using ANOVA to gain insight into specific measures which are characteristic of CL. Of the time domain features, Hjorth complexity of the sloped signal had the highest effect size ($F(2,66) = 9.50$, $p < .01$, $\eta^2 = .13$). In the frequency domain, the mean frequency of the slope ($segpoints = 10$) was found to have the highest value ($F(2,66) = 13.88, p < .01, \eta^2 = .17$). Of the frequency domain features extracted from the PSD for $segpoints = 40$, the PSD coefficient at 0.065 Hz had the highest value ($F(2,66) = 9.53$, $p < .01$, $\eta^2 = .13$)). For comparison, the effect size of Accumulated GSR, which was previously found to be informative of CL [14], was also measured ($F(2,66) = 4.32$, $p = 0.017$, $\eta^2 = .061$)).

4.2 Classification Model

The F1-Scores from leave-one-subject-out validation during model selection were recorded for both days and are shown in Fig. 2. Note that only the results using the K-Nearest Neighbour algorithm for fitness evaluation is shown, since Naïve Bayes performed at the level of a random classifier. The graphs demonstrate the segpoints values that had the best F1-Score (and thus were chosen for the model on each day). These were segpoints = 40 on day one (F1-Score = 0.406) and segpoints = 30 on day two (F1-Score = 0.434).



(a) Day 1 Scores



(b) Day 2 Scores

Fig. 2. Comparison of F1-Scores for varying segpoints in model selection stage on day one (a) and day two (b) using genetic algorithm-based feature selection

The results from the model testing stage are shown in Fig. 3, comparing the performance of the models trained/ tested using K-Nearest Neighbours, Random Forest and SVM Linear, alongside the Random Baseline performance and a model generated using Accumulated GSR for comparison (with Random Forest as learning algorithm since it performed best in combination with Accumulated GSR). SVM Linear had the best performance across both tests. The model trained on day one and tested on day two had an F1-Score of 0.365, and the model trained on day two and tested on day one had an F1-Score of 0.3774.

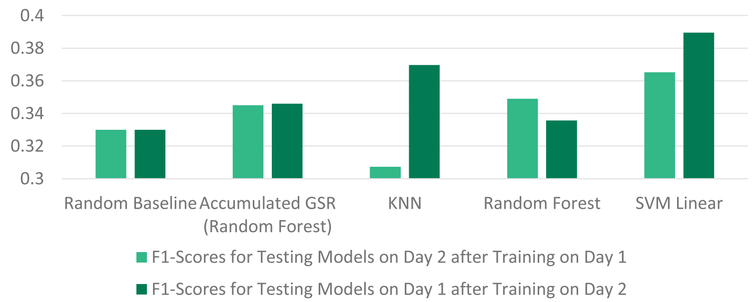


Fig. 3. Comparison of F1-Scores across different algorithms in model testing stage using genetic algorithm-based feature selection

5 Discussion

Our results for effect size during feature exploration indicate that Mean frequency of slope (segpoints = 10), PSD coefficients of slope at 0.065 Hz (segpoints = 40) and Hjorth complexity of slope were more discriminative between CL levels compared to Accumulated GSR, a feature used in previous work [14]. This suggests that, at least for our data, which is based on a multiple object tracking task as compared to solving arithmetic tasks, these new metrics may better characterise different CL levels.

When comparing segpoints, setting segpoints to 40 on day one maximised number of significant frequency domain features. Setting segpoints allows selection of trade-off between variance and frequency resolution of the PSD, so segpoints = 40 could represent a good trade-off for feature extraction from the PSD.

However, more significant features associated with a segpoints, does not necessarily mean that the features generated for the segpoints are more informative of CL. Note that even though segpoints = 40 had the greatest number of significant features, mean frequency for segpoints = 10 had the highest effect size. This suggests that there may be individual features for specific segpoints that are highly indicative of CL, even though they may contain fewer significant features. The obvious limitation of this analysis approach is that it does not take the effect of using multiple features as an indication of CL into consideration. This is addressed through the analysis of the classification models.

The model selection results from the genetic algorithm-based feature selection (Fig. 2) provide some insight into an appropriate segpoints value for generating the PSD coefficients. We found that $segpoints = 40$ and 30 were the best performing values in both day one and two. Note that the range of segpoints values tested was 10 to 60 with increments of 10, and so the model selection results suggests that a mid-range trade-off (given the sampling frequency of 20 Hz and recorded duration of 2 min) sets the best frame size and trade-off between frequency resolution and estimate variance. Such a result agrees with the analysis from the feature exploration section, where $segpoints = 40$ had the most number of significant features.

Model testing results (Fig. 3) indicate the effectiveness of the approach towards generating a model that generalises to unseen data. The classifiers (except for KNN) performed better than Random Baseline, and therefore seem to be capturing some of the effect (cognitive load influencing GSR). The inconsistency of KNN could be a consequence of being an inappropriate classification mechanism.

The results also show the SVM Linear models had the highest scores in both tests, suggesting it generalises the best. Furthermore, it consistently performed better than Accumulated GSR models, such that combining multiple relevant features using SVM Linear is better than solely relying on Accumulated GSR. However, performance still left much to be desired, with F1-scores slightly below 0.4 on three-class classification during testing. Perhaps the measured signal was externally influenced by factors aside from CL, or the features and classifier were not the most suitable choices for capturing the effect. Also, the features were not optimised for the classifier (due to computational limitations), since KNN was used for feature selection and SVM Linear for training/testing.

6 Conclusions

In this study we have conducted an investigation into classification techniques for human cognitive workload using Galvanic Skin Response, when performing a modified multiple object tracking task, *UAV Defender*, over two consecutive days of testing. The current methodology has been shown to outperform previously used metrics, and provide moderate discriminability for CL between levels of task difficulty, although the effects described herein leave room for improvement. These results are encouraging and provide justification for further research, including implementing the current methodology on new data sets, as well as testing other GSR devices (as there is some question of fidelity with the low-cost GSR sensors employed here). While affordability is a key aspect of our proposed system, we need to maximise the quality of the original GSR signal to promote the most accurate level of classification possible.

One of the goals of this research is to develop *lightweight classification systems* for use in critical task environments, where traditional means of workload detection, and more complex psychophysiology-based measures are unsuitable (e.g. environments such as road transportation, maritime transportation, operation of heavy machinery, etc.). We believe that the combination of low-cost equipment, reliable and relatively non-invasive sensors, and sophisticated data processing techniques will allow us to monitor CL in critical task environments as well as in the rising field of human-in-the-loop autonomous systems, where monitoring of autonomous systems is necessary.

Although a real-time system could be designed based on our analysis pipeline, there are some considerations that need to be taken into account. The feature extraction we used in this study occurred on signals using a 2-min time window. In practice, this would mean a long delay for changes in CL to be measured.

Further, due to the size of our sampling population, we applied within-subject standardisation in our analysis for both GSR signals and features. This implies that a deployed system would work in a user-dependent fashion, with every user having to perform a training phase. However, training the model with a larger pool of participants is likely to generate generalisable models that would perform well on new target users. This will be part of future work.

Acknowledgements. This research was funded in part by the CSIRO Data61 Automation Trust and Workload CRP, and Australian Research Council DP160101891, CERA247.

References

1. Ayaz, H., et al.: Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In: Rice, V. (ed.) Advances in Understanding Human Performance, vol. 20105280, pp. 21–31. CRC Press, Boca Raton (2010)
2. Baddeley, A.D.: Working memory. Science **255**(556–559), 5044 (1992)
3. Cain, B.: A review of the mental workload literature. Technical report, DTIC (2007). http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA474193
4. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**(Jul), 2079–2107 (2010)
5. Dünser, A., Mancero, G.: The use of depth in change detection and multiple object tracking. Ergon. Open J. **2**, 142–149 (2009)
6. Bureau d'Enquetes et d'Analyses France (BEA): Final Report: Accident to Airbus A330–203 Registered F-GZCP, Air France AF 447 Rio de Janeiro - Paris, 1st June 2009. Technical report, Air Accident Investigation Unit (AAIU), October 2014. http://www.aaiu.ie/node/687
7. Funke, G., et al.: Evaluation of subjective and EEG-based measures of mental workload. In: Stephanidis, C. (ed.) HCI 2013. CCIS, vol. 373, pp. 412–416. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39473-7_82
8. Healey, J., Picard, R.: SmartCar: detecting driver stress. In: Proceedings 15th International Conference on Pattern Recognition, ICPR-2000, vol. 4, pp. 218–221 (2000)
9. Hjorth, B.: EEG analysis based on time domain properties. Electroencephalogr. Clin. Neurophysiol. **29**(3), 306–310 (1970)

10. International Organization for Standardization: ISO 17488:2016(en), Road vehicles - transport information and control systems - detection-response task (DRT) for assessing attentional effects of cognitive load in driving. Technical Report ISO 17488:2016(en), International (2016)
11. Lim, C.L., et al.: Decomposing skin conductance into tonic and phasic components. Int. J. Psychophysiol. **25**(2), 97–109 (1997)
12. Lochner, M., Duenser, A., Lutzhoft, M., Brooks, B., Rozado, D.: Analysis of maritime team workload and communication dynamics in standard and emergency scenarios. J. Shipp. Trade **3**(1), 2 (2018)
13. Lochner, M.J., Trick, L.M.: Multiple-object tracking while driving: the multiple-vehicle tracking task. Atten. Percept. Psychophys. **76**, 2326–2345 (2014)
14. Nourbakhsh, N., Wang, Y., Chen, F.: GSR and blink features for cognitive load classification. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013. LNCS, vol. 8117, pp. 159–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40483-2_11
15. Phinyomark, A., Thongpanja, S., Hu, H., Phukpattaranont, P., Limsakul, C.: The usefulness of mean and median frequencies in electromyography analysis. In: Naik, G.R. (ed.) Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges. INTECH Open Access Publisher (2012). https://doi.org/10.5772/50639
16. Posada-Quintero, H.F., Chon, K.H.: Frequency-domain electrodermal activity index of sympathetic function. In: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 497–500, Febraury 2016
17. Pylyshyn, Z.W., Storm, R.W.: Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat. Vis. **3**(3), 179–197 (1988)
18. Rozado, D., Duenser, A.: Combining EEG with pupillometry to improve cognitive workload detection. IEEE Comput. **48**(10), 18–25 (2015)
19. Rozado, D., Lochner, M., Engelke, U., Duenser, A.: Detecting intention through motor-imagery-triggered pupil dilations. Hum.-Comput. Inter., 1–31, Febraury 2017
20. Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F.: Galvanic skin response (GSR) as an index of cognitive load. In: CHI 2007 Extended Abstracts, pp. 2651–2656. ACM, New York, April 2007
21. Uguz, H.: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl.-Based Syst. **24**(7), 1024–1032 (2011)
22. Welch, P.: The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967)
23. Wilson, G.F., Fullenkamp, P., Davis, I.: Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. Aviat. Space Environ. Med. **65**(2), 100–105 (1994)
24. Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A., Li, Z.: Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans. Comput.-Hum. Interact. **21**(6), 1–23 (2015)