Using biometric measurements to compare graphical user interfaces

Dmitriy A. Kostiuk, Oleg O. Latiy, Anastasia A. Markina, Vadim P. Shamonin
Brest State Technical University, 267, Moskovskaya Str., Brest, Belarus, 224017,
dmitriykostiuk@gmail.com, http://en.bstu.by/

Abstract: *The analysis of the approaches to evaluate human interaction with GUI is carried out. Usage of the contemporary consumer electronic devices with biometric sensors as a human body monitoring equipment is substantiated. Testing schemes are proposed based on the examination of the operator's interaction with a graphical environment. Set of criteria for biometric evaluation of the human-computer interaction effectiveness is formulated.*

Keywords: GUI, ergonomics, biometric sensors.

1. INTRODUCTION

Approaches and methods to evaluate the effectiveness of user's work with software can be divided into two different groups. Classical methods heavily rely on the participation of an ergonomics expert, who makes judgments based on interviews, timing measurements and video-logging of how do users perform specific test tasks while working with the compared software products [1]. This approach requires a lot of time for the analysis, and the quality depends a lot on the expert's professional skills.

Approaches and methods of the other type are based on the idea to monitor user's physical and mental state with a set of special measuring devices. Registration of biometrical parameters related to cognitive and physical loads, such as heart rate, galvanic skin response, brain waves and others, allows to define bottlenecks in human-computer interaction much faster and have weaker dependence on expert's judgment. As a consequence, instrumental evaluation allows to formulate a set of proposals to improve software faster [2, 3].

Of course, biometric measurements are not able to provide a clear picture of the advantages of one GUI over another. After series of experiments, the researcher gets a big amount of numerical data that requires processing and interpretation no less than what the survey and logging materials obtained with the classical approach. The difference is in the possibility of their automatic interpretation according to predefined computational criteria.

Until recently, this approach was limited due to the low prevalence and high cost of the required equipment. However, lately a significant number of devices with biometric sensors have appeared in the area of fitness and entertainment. An approach to use biometric capabilities of such popular devices combined with a set of simple criteria targeted at a comparative analysis of the software ergonomics is presented in this work.

2. CHOICE OF THE MEASURED PARAMETERS

The list of sensing technologies used in consumer-grade biometric devices includes, first of all, photoplethysmographic pulse measurement, which was first developed for sport heart rate monitors, and then adopted by fitness trackers and smartwatches (fig. 1). Some devices can measure the electrical conductivity of skin (fig. 2) (for auxiliary purposes), which depends on skin moisture, provided by sweat glands controlled by the sympathetic nervous system [4, 5]. For this reason, skin conductivity is often used as an additional indicator of psychological or physiological arousal used in the pair measurements combined with a heart rate [6].

Also there are entertainment gadgets that can record brain activity to determine the concentration of the user's attention. Such devices have few on-skin sensors to measure the general level of the brain neurons electrical activity, and can evaluate not much parameters beside the attention level (by measuring EEG beta frequencies), but that's exactly what is needed for the comparison.

Finally we should mention one more capability of the typical fitness trackers, i.e. measurements of the user's kinematic activity. Typically such devices use inertial systems to determine the acceleration of an object and its angular velocities by using of an accelerometer and a gyroscope with further analysis of the displacement. While such approach is usable for fitness, it's not too good at distinguishing the physical activity of the non-entertainment human-computer interaction, such as typing and operating the mouse pointer. An absolutely different type devices can be recommended for such measurements – ones, recording the electrical activity of muscles. Usage of non-invasive electrodes put on skin resembles skin conductivity measurements for some extent, but records bioelectric potentials arising in skeletal muscles during the excitation of muscle fibers (myography). Being combined with the pattern recognition software such measurements allow tracking the movements of fingers – the task fitness trackers can not be used for (fig. 3). It should be noticed that myography is the most experimental technology in this list of biometric measurements, but still there are several projects targeted at bringing it to the mass market.

All these devices are able to carry on continuous monitoring, can transfer data to a personal computer and at the same time, due to mass production, are widely available.

During the approbation, the following biometric parameters were selected for monitoring: electrodermal activity of the skin (EDA), electroencephalogram (EEG) rhythms and heart rate (HR), as far as EEG waves recorded by an electroencephalograph [3, 7].

3. TEST TASKS SPECIFICS

We have researched two variants of user interaction with a graphical interface series of different type operations in one program (i) and a long sequence of routine operations (ii).

While performing a series of different type operations in one program, the user is supplied with a set of tasks under one general thematic direction (e.g. the work with any large document in a word processor). In this case, the

research is designed to assess how the overall layout and dynamics of the application interface affect the user, and, in particular, how much the toolbars are adapted for the actions [4]. This set of tasks is relevant when comparing the convenience of several application programs in the same subject area.

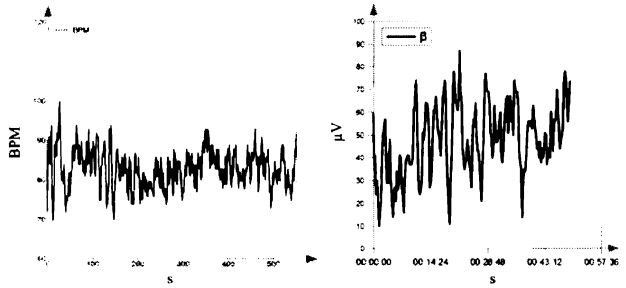

Fig.1 – Time series of HR (on the left) and EEG beta waves (on the right)

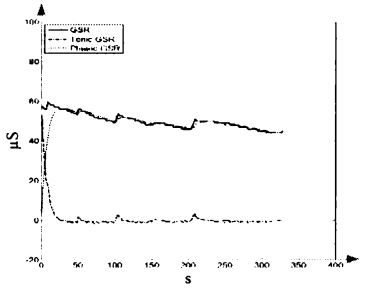

Fig.2 – Electrodermal activity time series

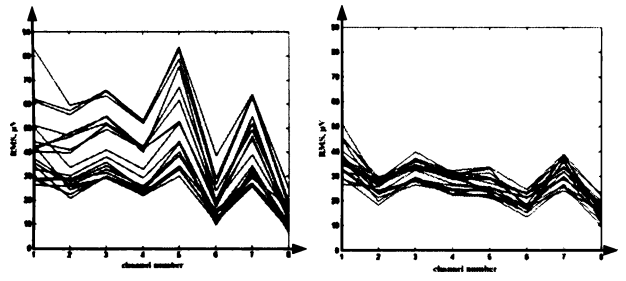

Fig.3 – Electrical activity of muscle fibers used for the fingers movements recognition

Performing a long sequence of routine operations. Involves large amount of tasks, all of which are of the same type; each task involves several applications, or several parts (windows, frames, panels) of the same application. The primary task in this case is to evaluate the contribution of the GUI auxiliary elements: task bars, window controls, cursor positioning features, system-wide widgets (for example, context menus), etc. This is an assessment of the impact caused by the operating system's graphical shell on the fatigue caused by monotonous work and the mind concentration of the user. Tests can be used with standard applications (e.g. typical operations in the file manager), or with applications developed specifically for the test.

Approbation of the testing methodology for the first version of the interaction was carried out on ergonomics assessment of office packages. A comparison was made between interfaces with a Microsoft Fluent Interface / ribbon toolbar, a classic top panel and a side panel. We have tested the Word and Excel applications from the Microsoft Office 2007 package, and Writer and Calc from LibreOffice 5.0 in two different modes of the interface. During the experiment, users where performing formatting, changing text and tables markup [8, 9].

Research of the second version of human-computer interaction was performed during testing the window management of graphical shells in modern Unix-like systems (KDE Plasma Desktop with the task bar at the bottom of the screen, Gnome 3 with a special view mode for switching windows and the Unity shell from the Ubuntu Linux distribution with the dock-panel used to switch windows). Choice of routine operations was made for copy/paste of the text fragments using the context menu, as well as memorizing the geometric figure and its search among 25 different variants. The user's work was associated with a lot of windows switching, as a result of which the graphic shell acted as an external load [10, 11].

4. EVALUATION CRITERIA

We introduced five user's performance indicators: the given actions duration τ, the number of errors e, the heart rate p, the attention concentration β, and the skin electrical conductivity change g. Providing penalty functions q_i for each of these components to reflecting their importance in terms of the expected result, we can formulate a general criterion for the quality of work:

$$Q = \tau \cdot q_t + e \cdot q_e + p \cdot q_p + \beta \cdot q_\beta + g \cdot q_g. \quad (1)$$

Errors made by the operator when performing atomic operations are unequal, and to account for them, the parameter e is divided into the missing errors $e^{(1)}$ and the corrected errors $e^{(2)}$:

$$e = e^{(1)} + e^{(2)}. \quad (2)$$

The component $e^{(2)}$ increases the work execution time and, thus, does not require additional accounting, and the component $e^{(1)}$ becomes an independent indicator in formula (1). Also in some cases, user's failure affects several operations in sequence. For this reason, the duration of the failure τ_e is also informative (e.g., the average value during the test run).

We have chosen the rate of tasks completion υ as an indicator of the speed of work:

$$\upsilon = \frac{\partial S}{\partial \tau} = \sum_i \frac{[1 - \delta(e^{(2)} - e_i^{(1)})]S_i}{\tau_i}, \quad (3)$$

where S_i is the i-th atomic operation (e.g., it can be one printed character, line shifting, etc.), $e_i^{(2)}$ is a sign of the error presence in the execution of the i-th atomic operation, and δ is the Dirac delta function.

Mostly, the experiments were estimated by the average rate equal to the number of correctly performed atomic operations generated by the operator per second:

$$\upsilon = [S - q_e \cdot e^{(1)}]/\tau, \quad (3')$$

where S is the total number of performed operations, $e^{(2)}$ is the number of operations that were not performed, or performed with an error. The penalty function q_e can be used to compare the quality of work: with this function, the "weight" of mistakes increases in proportion to their importance for a particular task being solved. If the degree of importance of error-free work is not defined the value if

q_e = 1 is used, in which <υ> is equal to the number of correctly performed operations per second.

The mean HR during the test <p> and the mean value of the concentration of attention <β> where chosen as an informative parameter in assessing the HR and concentration of attention. In practice we have evaluated mind concentration by the "Attention" metric of the Neurosky Mindwave encephalograph, associated with the β-rhythm of the brain [12, 13].

High-frequency phasic g_p and low-frequency tonic g_t galvanic skin response (GSR) are identified in EDA:

$$g = g_p + g_t. \quad (4)$$

Phasic GSR has the form of short-duration pulses in response to external stimuli or to anxiety, tension, and thought activity. Tonic GSR serves as an indicator of a person's functional state and, according to existing researches, is a less universal indicator: it is a slowly changing component, and its recording requires calibration for each user [14]. Therefore, in the estimation of GSR, the data obtained from the EDA sensors must be filtered to allocate the phasic component, on the basis of which the number of extremums g_p can be counted:

$$g_{pi} = \begin{cases} 1, & g'_p(t_i) = 0, \ g''_p(t_i) < 0; \\ 0, & otherwise. \end{cases} \quad (5)$$

5. CONCLUSION

Thus, the list of biometric indicators proposed for the practical evaluation of the operator's performance includes the following parameters: the rate of the actions υ, the number of errors missed by the operator e_1, the duration of the failure τ_e, the HR p, the attention level of the operator β, and the phasic GSR g_p. Based on the test results, the average values for the listed parameters, and, additionally, the maximum deviation of the parameter from the average component are calculated. Each of the presented criteria allows to create the time series containing the values measured during the test (HR, EDA, etc.) or returned by the testing program (testing time and errors) to a single value that reflects the nature of the work of a specific user in a particular test. Optionally, in case of presence of weight coefficients reflecting the importance of each indicator in the operator's work, the calculated values can be reduced to a common integral estimate in expression (1).

A simple comparison of the values obtained during the testing the user's work with each of the compared products can be performed to compare several software products. As a result, the program that showed the best values for the majority of users tested is considered more appropriate.

6. REFERENCES

[1] Zabrodin U.M. Metodologicheskie problemy funktsionalnogo sostoyaniya cheloveka-operatora // Voprosy kibernetiki. Psihicheskie sostoyaniya i effektivnost deyatelnosti. M.: 1983. – S. 3–25

[2] Zhuravskiy V.I., Kostiuk D.A., Latiy O.O., Markina A.A. Programmno-Apparatnaya sistema dlya sravnitelnyh issledovanij ergonomiki programmnogo obespecheniya // Informacionnye tehnologii i sistemy 2015 (ITS 2015): Materialy mezhdunarodnoy nauchnoy konferentsii. Minsk, BGUIR, 29 oktyabrya 2015 g. – S. 252–253.

[3] Rebsamen B., Kwok K., Penney T.B. Evaluation of cognitive workload from EEG during a mental arithmetic task // Proceedings of the human factors and ergonomics society annual meeting. Vol. 55, Iss. 1, 2011 . – P. 1342–1345.

[4] Raskin Dzh. Interfeys: novye napravleniya v proektirovanii kompyuternyh sistem. - SPb.: Simvol-Plyus, 2003. – 272 S.

[5] Martini F., Bartholomew E. Essentials of Anatomy & Physiology // San Francisco: Benjamin Cummings, 2003. – P. 267.

[6] Chen W. Continuous estimation of systolic blood pressure using the pulse arrival time and intermittent calibration // Medical and Biological Engineering and Computing. – Vol. 38, 2000. – P. 569–574.

[7] Guselnikov V. I. Elektrofiziologiya golovnogo mozga. - M: Vysshaya shkola, 1976.

[8] Kostiuk D.A., Latiy O.O., Markina A.A. Instrumentalnaya otsenka sostoyaniya polzovatelya v zadache sravneniya interfeysov ofisnyh prilozheniy // XII konferentsiya razrabotchikov svobodnyh programm. Tezisy dokladov. – Kaluga, 16-18 oktyabrya 2015 g. – M.: Alt Linuks, 2015. – S. 8–12.

[9] Kostyuk D.A., Latiy O.O., Markina A.A. Ob effektivnosti ispolzovaniya metafory lentochnogo interfeysa // Odinnadcataya konferenciya «Svobodnoe programmnoe obespechenie v vysshej shkole»: Materialy konferentsii. – Pereslavl, 30-31 yanvarya 2016 g. – M.: Alt Linuks, 2016. – S.17–23.

[10] Kostyuk D.A., Kostyuk K.L., Derechennik S.S., Tavonius K.A., Shitikov A.V. Issledovanie effektivnosti pereklyucheniya okon v sovremennyh graficheskih interfeysah // Vestnik BrGTU. 2011. № 5 (71): Fizika, matematika, informatika. – S. 45–48.

[11] Kostiuk D.A., Derechennik S.S., Shitikov A.V., Latiy O.O. Approach to evaluate effectiveness of human-computer interaction with contemporary GUI // Tretya mizhnarodna naukovo-praktichna konferenciya FOSS Lviv 2013: zbirnik naukovyh prac, Lviv, 18–21 kvitnya 2013 r. – Lviv, 2013. – S. 85–87.

[12] Dhali S. A Study of Brainwave eSensing Activity. Department of Computer Science, Malmo University (electronic publication). https://www.overleaf.com/articles/bci/mcsvkjwhcffb/viewer.pdf

[13] Sezer A., Inel Y., Seçkin A.Ç., Uluçinar U. An Investigation of University Students' Attention Levels inReal Classroom Settings with NeuroSky's MindWave Mobile (EEG) Device. // Proc. of IETC 2015 int. conf., May 27-29, Istanbul, Turkey. – P. 88-101.

[14] Benedek M., Kaernbach C. A continuous measure of phasic electrodermal activity // Journal of Neuroscience Methods, №. 190, 2010. – P. 80–91.