2006

Interface: assessment of human-computer interaction by monitoring physiological and other data with a time-resolution of only a few seconds

K Hercegfi
hercegfi@erg.bme.hu

O. Kiss
kiss.orhidea@ppk.elte.hu

K. Bali
kriszta@erg.bme.hu

L. Izso
izsolajos@erg.bme.hu

INTERFACE: ASSESSMENT OF HUMAN-COMPUTER INTERACTION BY MONITORING PHYSIOLOGICAL AND OTHER DATA WITH A TIME-RESOLUTION OF ONLY A FEW SECONDS

Hercegfi, Károly, Budapest University of Technology and Economics, Egry J. u. 1, 1111 Budapest, Hungary, hercegfi@erg.bme.hu

Kiss, Orhidea Edith, Eötvös Loránd University, Kazinczy u. 23, 1075 Budapest, Hungary, kiss.orhidea@ppk.elte.hu

Bali, Krisztina, Budapest University of Technology and Economics, Egry J. u. 1, 1111 Budapest, Hungary, kriszta@erg.bme.hu

Izsó, Lajos, Budapest University of Technology and Economics, Egry J. u. 1, 1111 Budapest, Hungary, izsolajos@erg.bme.hu

Abstract

Earlier publications have shown that a Heart Period Variability (HPV) -based methodology, after careful adaptation, could be a powerful technique for monitoring mental effort in Human-Computer Interaction.

This paper outlines the INTERFACE testing workstation developed by researchers of our department. This system is based on the simultaneous assessment of HPV, time data of keystroke and mouse events, video images of users' behaviour and screen content, etc. It is capable of identifying quality attributes of software elements with a time-resolution of only a few seconds.

Our series of experiments demonstrate the practical usability of this improved methodology for testing user interfaces. The method of analysis allows us to decide what types of problems are significant to the users, and what types of problems set back the users only slightly. On the other hand, the method allows us to decide, to what extent the found problems and their assessed severity concern all the users in general, or how these things depend on the type and characteristics of the users.

At the end of this paper, we will give a brief description of the further development of this INTERFACE methodology: we are in the process of integrating also another physiological channel – Skin Conductance (SC).

Keywords: Usability Evaluation, Empirical Methods, Heart Period Variability (HPV).

1 INTRODUCTION

While many current usability testing systems collect various performance measures, very few attempt to measure cognitive effort. The basic advantage and novelty of the methodology proposed in this paper lies in its capability of recording continuous on-line data characterising the user's actual mental effort derived from Heart Period Variability (HPV) simultaneously and synchronised with other characteristics of HCI. This way a very detailed picture can be obtained which serves as a reliable basis for the deeper understanding and interpretation of mechanisms underlying HCI.

Elementary steps of Human-Computer Interaction (HCI), like the different mental actions of users followed by a series of keystrokes and mouse-clicks, are the basic and usually critical components of using information technology systems. These steps can be modelled and analysed by experts, but empirical studies of real users' interaction often highlight new HCI problems or give more objective results than expert analyses. One of the key aspects of empirical methods is measuring *mental effort* as it is laid down e.g. in the earlier international standard of software product evaluation (ISO/IEC 9126:1991). Hence we need methods capable of validly and reliably monitoring users' actual mental effort during these *elementary* steps.

To attain the above, a complex methodology was developed earlier at our department, by Prof. Lajos Izsó and his team (Izsó 2001, Izsó & Láng 2000, Izsó & al. 1999, Izsó & Ziljstra 1999). This paper intends to provide an improved methodology with a case study.

The structure of this paper is as follows:
- The description of the INTERFACE methodology, with special respect to the heart period variance (HPV) parameter as opposed to other alternative physiological measures. An emphasis is given to the fine grain temporal resolution that has recently been achieved.
- Illustrating case studies are provided.
- A brief description of the further development of this INTERFACE methodology.
- Summary.

2 THE INTERFACE METHODOLOGY AND WORKSTATION

Figure 1 shows the conceptual arrangement of the INTERFACE (INTegrated Evaluation and Research Facilities for Assessing Computer-users' Efficiency) workstation.

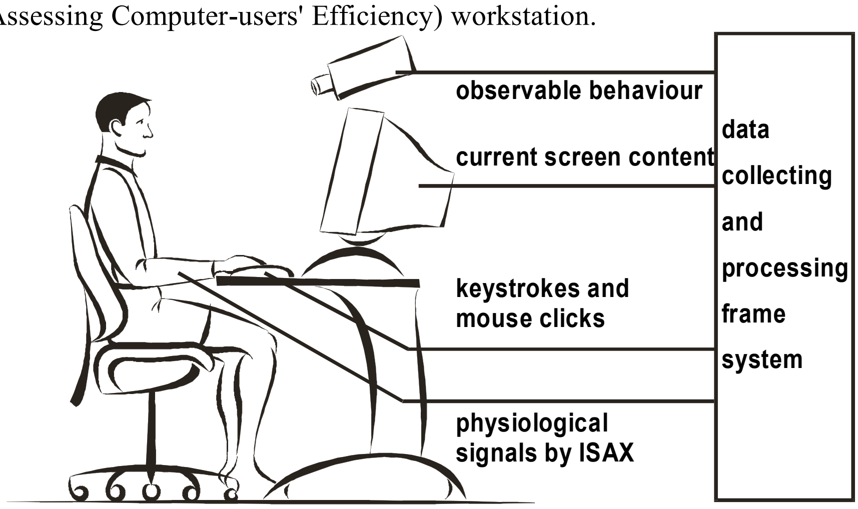

Figure 1. Conceptual arrangement of the INTERFACE user interface testing workstation.

The method simultaneously investigates the following:

- Users' observable actions and behaviour
    - Recording the keystroke and mouse events.
    - Video recording of the current screen content.
    - Video recording of users' behaviour (acts, mimics, gestures etc.) during working with the given software interface.
- Special psycho-physiological parameters
    - One of them is the power spectrum of Heart Period Variance (HPV), which is regarded as an objective measure of current mental effort and a reliable indicator of user state.
    - At the end of this paper, we will describe the further development of this methodology: we are in the process of integrating also another physiological channel – Skin Conductance (SC).

In addition to observable elements of behaviour, the proposed complex method also includes more traditional questionnaires and interviews assessing mental models, subjective feelings and opinions of users about their perceived task difficulty and fatigue.

The basic advantage and novelty of this methodology lies in its capability of recording continuous on-line data characterising the user's actual mental effort simultaneously and synchronised with other characteristics of HCI. This way a very detailed picture can be obtained which serves as a reliable basis for the deeper understanding and interpretation of mechanisms underlying HCI. Recording these various data simultaneously requires a more sophisticated technical background than other empirical methods based on only personal observation or video recording. However, the application of this methodology is not much more complicated. It can be used to make traditional interviews and video analyses shorter and easier: synchronised data are available 10 minutes after recording, and the researcher can focus on replaying periods of high mental effort. Multiple channels enable researchers to concentrate on the channels that highlight the importance of various parts of the current event flow.

Naturally, it may not be necessary to evaluate all types of software products with this complex empirical method. However, in certain cases, it is worth investing in a deeper analysis to provide more detailed results or to present results in a more effective way than simpler methods are capable of. This methodology is not standardized yet, but it can be a useful tool already.

2.1 Assessing Users' Performance and Behaviour

Performance measures are generally numbers of countable things performed during a period of time such as clients served, forms filled in, characters typed in, and items entered into a register or monthly returns, etc. Performance criteria often closely relate to time, in these cases certain predefined amount of work must be done during a given time unit. While measures mentioned above could be considered as measures of quantity, performance measures based on the number and nature of errors committed could be taken as indicators of the quality of computer work.

Error analysis is essential in studying HCI. In practice, however, it is often difficult to distinguish between accidental errors resulting e.g. from the differences of the preliminary knowledge of users and systematic errors attributable to design failures without an appropriate statistical analysis. A statistical analysis of the distribution of observed errors along the steps of a carefully chosen standard series of tasks could add further useful information to the obvious results of the study. For this purpose Prof. Lajos Izsó (2000) proposed a simple binomial model, which proved to be practically applicable.

Recording users' behaviour has outstanding importance. The video recording of the user's face and activity is an extremely rich source of psychological information as it directly reflects the mental state (e.g. boredom, routine activity in familiar environment, attention-demanding task, helplessness, getting lost, emotions like frustration, anger, joy, a laugh, etc.). Together with performance measures, the psychological analysis of video recordings makes it possible to identify the relative weak points of the particular software interface.

2.2 Assessing Mental Effort via Analysing Users' HPV Power Spectrum

A number of studies (Chen & Vertegaal 2004, Izsó 2001, Izsó & Láng 2000, Láng & Szilágyi 1991, Luczak & Laurig 1973, Mulder & Mulder-Hajonides van der Meulen 1973, Rowe & al. 1998, Sayers 1971) have shown that increasing mental load causes a decrease in the so-called mid-frequency (MF) peak of the heart period variance (HPV) power spectrum. To assess the spectral components of HPV power spectra in field studies, an integrated system called ISAX (Integrated System for Ambulatory Cardio-respiratory data acquisition and Spectral analysis) was developed and successfully used by Dr. Eszter Láng and her team. This equipment and method have been integrated into our system.

HPV proved to be an applicable, valid and robust indicator of mental effort in HCI (Izsó 2001, Izsó & Láng 2000, Izsó & al. 1999, Rowe & al. 1998). *Although there are other techniques for this purpose, either they are more difficult to evaluate and more invasive* (e.g. the EEG)*, or they give an overall, averaged indicator for a relatively long period of time*, from minutes to hours (e.g. the visual critical flicker frequency (CFF) and the practical realisations of bio-chemical measures). The main advantage of our method over the previously existing HPV-based methods is that the MF component of HPV shows changes in mental load *in the range of several seconds* (as opposed to the earlier methods with a resolution of ten seconds). It can be seen in the examples of the case study given below. This advantageous feature was achieved by an appropriate windowing data processing technique.

The method with the spectral analysis and the MF power are practically insensitive to the minor movements typically occurring in a sedentary job. The artefacts resulting from bigger movements (e.g. stretching) can be filtered out based on the video recording.

In our experience, participants forget about the electrodes and wires attached to their chest quickly. For example, during an on-site measurement at a telecom company, one of the phone operators who wore them stood up at the 5th minute of the recording, and tried to go to her colleague in the other corner of the office to ask something, forgetting about the additional wire between her and our computer. (This wire has since been eliminated.)

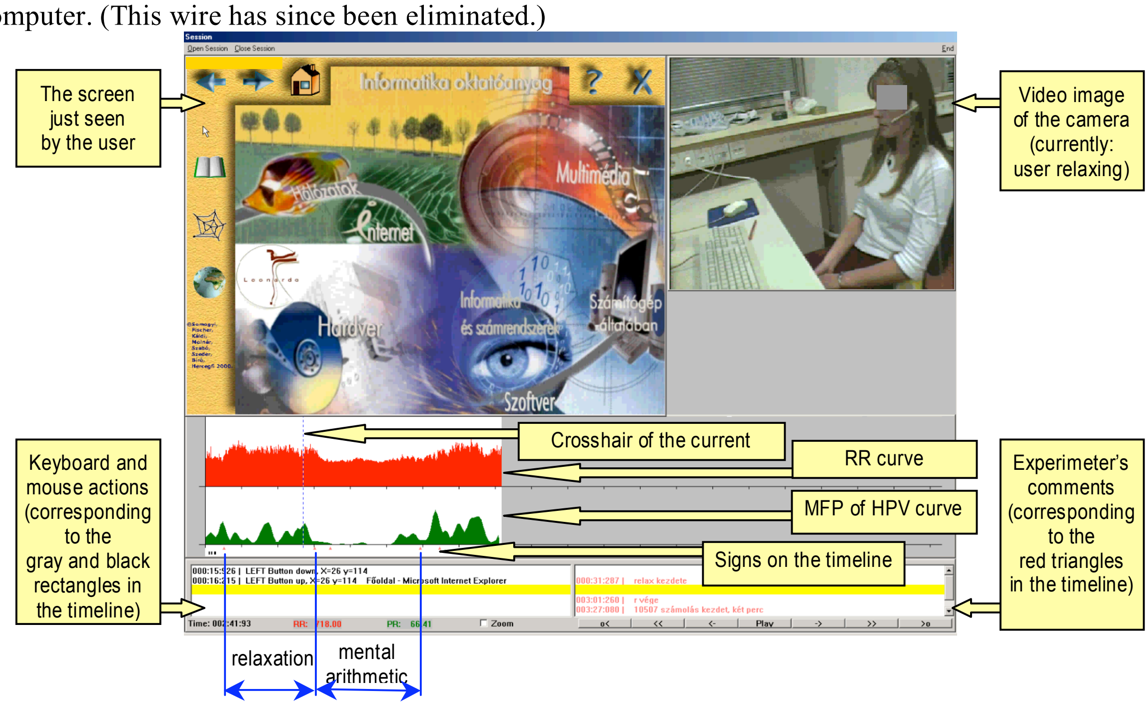

Figure 2. *The INTERFACE Viewer screen – a user during relaxation; the Mid-Frequency Power (MFP) of Heart Period Variance (HPV) shows relatively high values during relaxation and near-zero values during mental arithmetic; immediately after the mental arithmetic the rebound effect can be seen clearly.*

A typical INTERFACE result screen of a relaxing user can be seen in Figure 2. The parts of the INTERFACE Viewer screen and the typical change of the heart rate variability between relaxation and high mental effort can be seen in this figure.

3 RESULTS FROM A NEW AREA: APPLYING INTERFACE TO MULTIMEDIA SYSTEM EVALUATION

We have used INTERFACE in various areas (e.g. mail systems, a directory assistant service, an architectural CAD, WAP-based software). However, commercial sensitivities prevent publication of most of our findings. Therefore the following study is from a multimedia development project led by us. (NB Some minor pieces of our results from real industrial applications are available: Izsó (2001) and Izsó & Láng (2000).)

The case study presented in this section is from a quality assessment of a multimedia teaching package titled "Basics of Information Technology" (in Hungarian). It was developed in the framework of the Leonardo da Vinci Project "Developing and introducing multimedia teaching materials for vocational education" supported by the European Community. Károly Hercegfi, one of the authors, led the developer team; therefore he was thoroughly familiar with the concepts, the particular design elements and the planned HCI.

3.1 Methodology

The participants of the study were 24 students of two vocational secondary schools, who were performing learning sessions fitting to their studies with the help of the multimedia material. (3 subjects were participants of a pilot study with the same scenario as the scenario of the more thoroughly analysed series of experiments in which the other 21 subjects participated.) The students were informed that certain data were going to be recorded on them and were asked to allow placing ECG electrodes on their chest for recording HPV data with the data-collecting module of ISAX.

The 21 participants of the main series of experiments were 14 to 16 year old (9th graders). 13 of them studied at a vocational secondary school of economics, 8 of them studied at a technical vocational school. 14 of them were girls, 7 of them were boys. This combination was selected because this educational material – in spite of the technical aspects of its topic – was designed for 9th graders, whose subject called "Information Technology" is a general subject for all types of schools, not a special subject for only IT-oriented vocational schools.

Each session had the following schedule:

1. Collecting data about the current user:
   - filling in a questionnaire with demographic data, academic records, familiarity with the computer and the Internet, etc.;
   - MBTI (Myers-Briggs Type Indicator) psychological test to identify the cognitive style of the user. (The psychologists of our department were working on the localised adaptation of this test.)
2. 2-3 minutes' relaxation followed by a 2-3-minute period of mental arithmetic for "calibrating" curves depending on the heart rate variance.
3. Free browsing of the selected material for 5 minutes in order to get familiar with the style and controls of the CD.
4. Actual learning task: finding and learning two short sub-chapters in 10 minutes followed by a 2-minute knowledge-acquisition test. This part of the session aimed to have the students practise the typical usage of the material.

5. Searching task: finding short answers to 11 questions (20-25 minutes). This part of the session was the most important part for us. The sequence of the questions aimed to lead the users to the situations where various possible usability problems may occur.

6. Interview supported by playback (5 minutes).

The method of analysis allows us to decide *what types of problems* are significant to the users, and what types of problems set back the users only slightly. On the other hand, the method allows us to decide, to what extent the found problems and their assessed severity concern all the users in general, or *how* these things *depend on the type and characteristics of the users*.

We were able to focus to the quality attributes of software *elements* with a *time-resolution of only a few seconds.*

We were able to analyse the correlations of the rich data set gained by the INTERFACE system (e.g. time data, number of clicks, recorded tracking data in the hypermedia space, etc.) together with data obtained from the questionnaires and psychological tests of the 1st step mentioned above.

To execute the statistical calculations, we used the SPSS 12.0 for Windows software package.

A thorough empirical study with 21 subjects and 225 variables needs a lot of time, so there was not a realistic chance of increasing the number of subjects. Naturally, the series of experiments with only 21 subjects cannot be fully representative, but the selection of subjects succeeded in being varied enough:

- Previous experiences:
  - 9 students had computers for their personal use (not only for family use), 12 students did not.
  - The distribution of the levels of previous computer and Internet experiences were close to the normal distribution.
- Arts or sciences (technical, IT) interest:
  - 10 participants read literature regularly, 11 did not.
  - 7 participants read IT books and/or magazines, 14 did not.
  - In the list of the favourite subjects, 7 students listed IT, 8 students listed maths, 12 students listed arts subjects (e. g. literature, history, etc.).
  - The distribution of the maths grades was close to the normal distribution.
- General academic records:
  - The distribution of the general grade averages of the students was close to the normal distribution.
- Cognitive style:
  - The MBTI tests showed various types of dominant cognitive functions.

3.2 Examples of user interface problems identified with the INTERFACE system

3.2.1 A usability problems which we originally intended to focus on

A recurring problem was that some objects the users expected to be hot links were not clickable. The users got confused, because earlier during the same multimedia session they encountered certain graphic objects that actually were hot links, whereas the related texts were not. Figure 3 and Figure 4 show examples of the opposite: here the objects (bullets and images, respectively) were logically expected to be hot links, but they were not. This inconsistency resulted in unnecessary loads on the users and a substantial waste of time.

The advantage of the high temporal resolution can clearly be seen e.g. in Figure 3: the fine time structure of the HPV profile during the three subsequent clicks gives well established basis for interpretation.

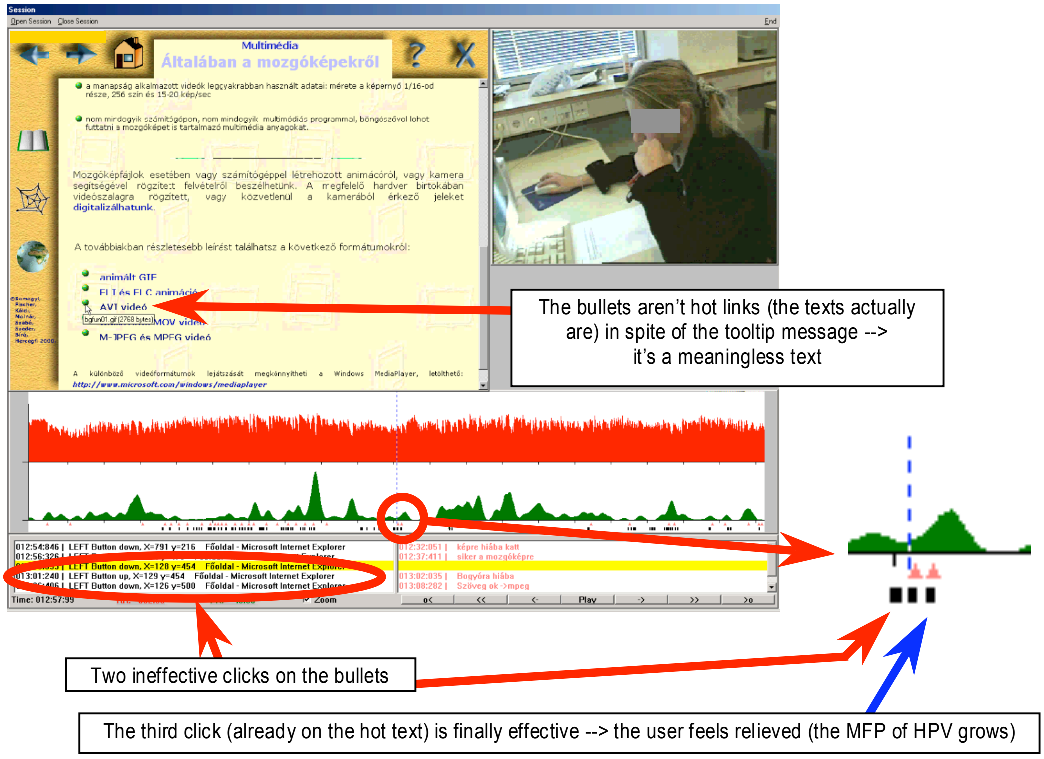

Figure 3. *The 6th user clicks on the bullets twice ineffectively, which resulted in a short period of unnecessary mental effort and losing valuable seconds.*

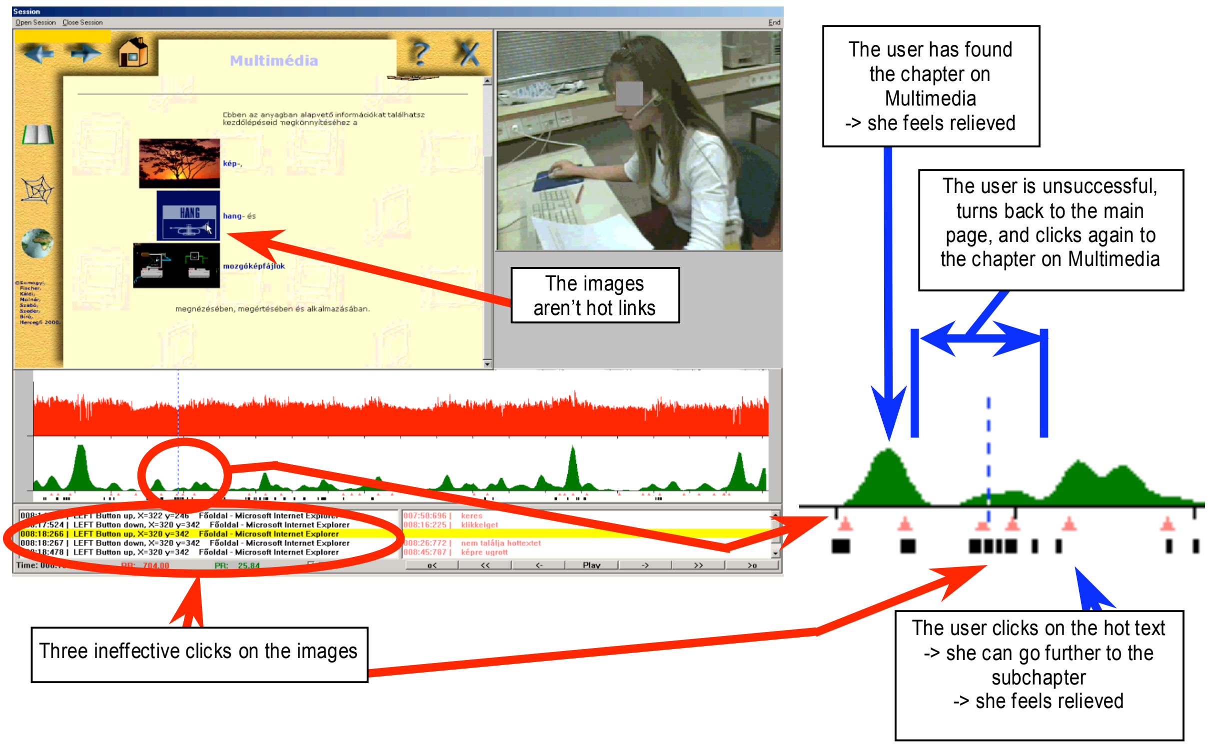

Figure 4. *The 2nd user clicks on the images two times ineffectively, which resulted in a short period of unnecessary mental effort and losing several seconds.*

An other example can be seen in Figure 4: the starting page of the chapter on Multimedia contains images that are not hot links. 71% of the participants clicked on the images first (ineffectively). They found the real hot links (the text instead of the images) after an average of 5.3 seconds' waste of time. The maximum delay was 80.5 sec.

The number of unnecessary clicks and the waste of time caused by this particular usability problem *do not correlate* with almost any of the other variables. (We have calculated Spearman rho values). The Kruskal-Wallis and Manny-Whitney statistical tests did not give any significant result. From practical point of view, this means that *this type of usability problem is a general problem; it affects all types of users*.

3.3 An unexpected usability problem

In the next two figures (Figure 5 and Figure 6), the difficulty of finding the scroll bar is demonstrated. The 6^{th} and 11^{th} users both discovered the scroll bar only after a helpless seven-minute trial-and-error searching. Figure 5 shows the 6^{th} user, the quite similar records of the 11^{th} user are not presented here. Although the 12^{th} user struggled on in this trap for only two minutes, her heavy mental efforts are clearly seen in the related MFP profile in Figure 6. (The origin of the problem was the following: the first part of the long, scrollable page – using this screen resolution – looks like a complete page; the figure and its caption are at the bottom of the current screen, as it can be seen at the Figure 5.)

The average waste of time caused by this problem was 69 seconds, with the maximum of 253 seconds.

But it is not these mere data that are most interesting. One third of the users did not have any problems here: they clicked on the scroll bar 1 to 3 seconds after they had arrived to this page. However, the other two thirds of the users needed 14 to 253 seconds. For example, the girl in Figure 5 had no less intellectual capacity or experiences with the Internet than the others. Why does this screen still represent a problem for her and the two thirds of the users, and why not for the others? *How does the severity of this usability problem depend on the type and characteristics of the users?*

- The Mann-Whitney statistical tests show the following:
  - The waste of time caused by this usability problem depends on the type of the school of the students: the students of the vocational school of economics wasted significantly more time here (p=0.006). However, in our case, all economics students were girls and most of the technical students were boys, so this effect cannot be separated from the effect of the gender: the significance level of the dependence on the gender is p=0.031.
  - The users who read literature regularly wasted more time than the others, with the significance level of p=0.021.
  - The users who read IT books and/or magazines regularly wasted less time than the others, with the significance level of p=0.013.
- The calculations of correlations – among others – show the following:
  - The students with better grade in maths wasted less time than the others; Spearman's rho is -0,441, with the significance level p=0.034.
  - ***The waste of time caused by this usability problem correlates strongly with scores on the Thinking–Feeling (T–F) dimension of the MBTI psychological test***; Spearman's rho is -0,533, with significance level of p=0.046. This result means, that **"thinking"-type users understood the logic** of the content and the user interface almost immediately, independently of the deceitful view of the screen; but **the users with "feeling" cognitive style had been fooled by** the apparent intactness of **the layout of the particular page**.

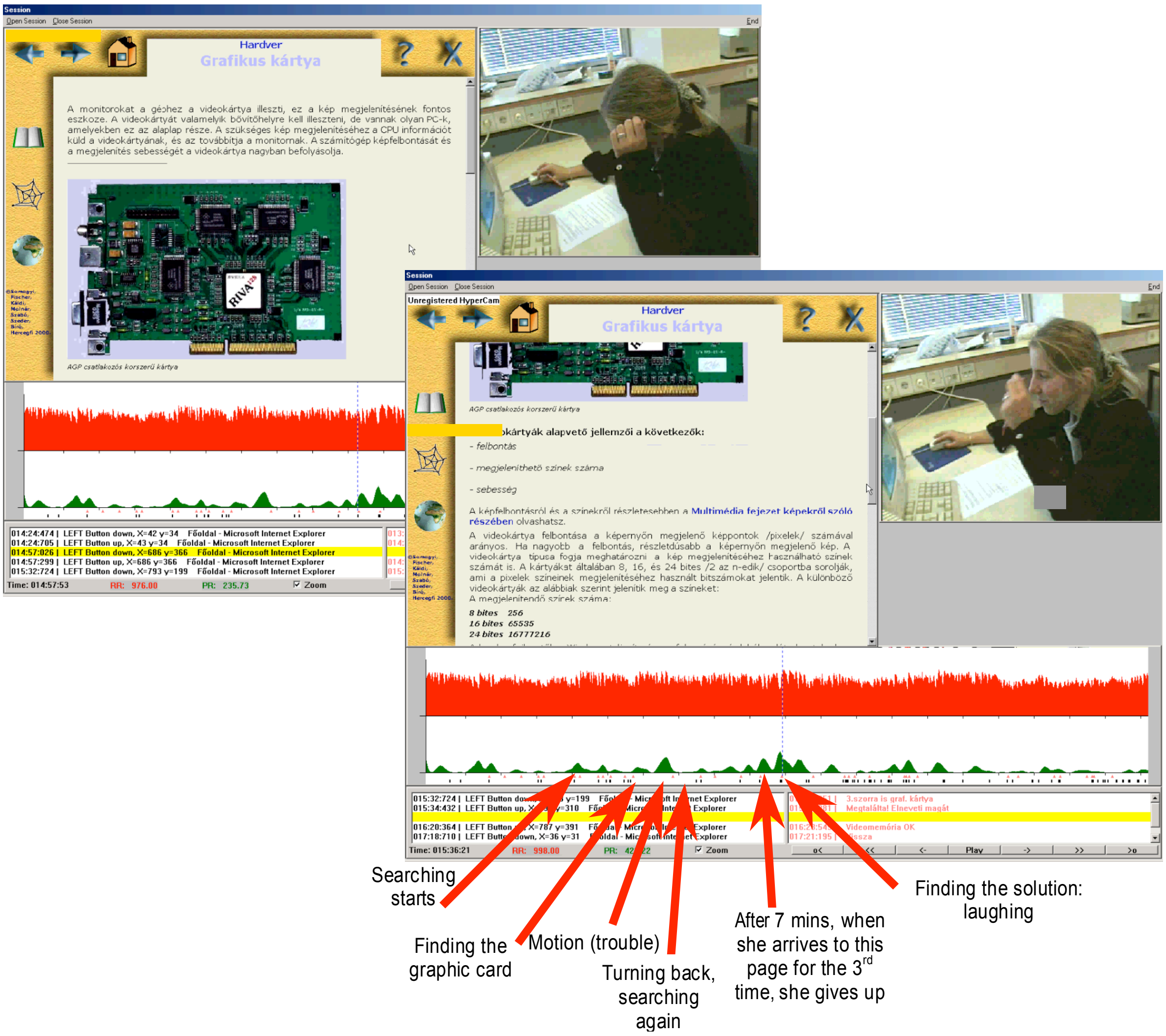

Figure 5. *After a 7-minute ordeal, the 6th user gives it up, but immediately after that, she finally discovers the solution (the scroll bar) and laughs. The upper screen shows a moment when the user is in the state of giving up, while the lower screen presents the situation a bit later when she just found the scroll bar.*

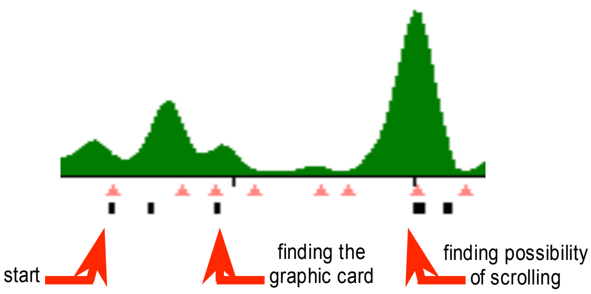

Figure 6. *The 12th user in the same situation as the 6th user in Figure 5. Although she struggled only for two minutes, her heavy mental investments are clearly seen in the much suppressed MFP profile during the period from finding the graphic card till finding the possibility of scrolling.*

4 FURTHER DEVELOPMENT OF THE INTERFACE METHODOLOGY AND WORKSTATION

In addition to the existing channels, we are in the process of integrating another physiological one: Skin Conductance (SC).

In contrast with our earlier experiments applying Heart Rate Variability (HPV) to monitor mental effort (Izsó 2001, Izsó & Láng 2000, Izsó & al. 1999, Izsó & Ziljstra 1999), measuring Skin Conductance in our INTERFACE methodology is new to us. We are working on it to complement our methodology with a component focusing on the *emotional* aspects of the HCI, in addition to our well-tried approach of *mental effort*.

We are at the very beginning of this development, but we can already present one of the first successful technical tests of the new version of the INTERFACE frame system integrating this new physiological channel (Figure 7).

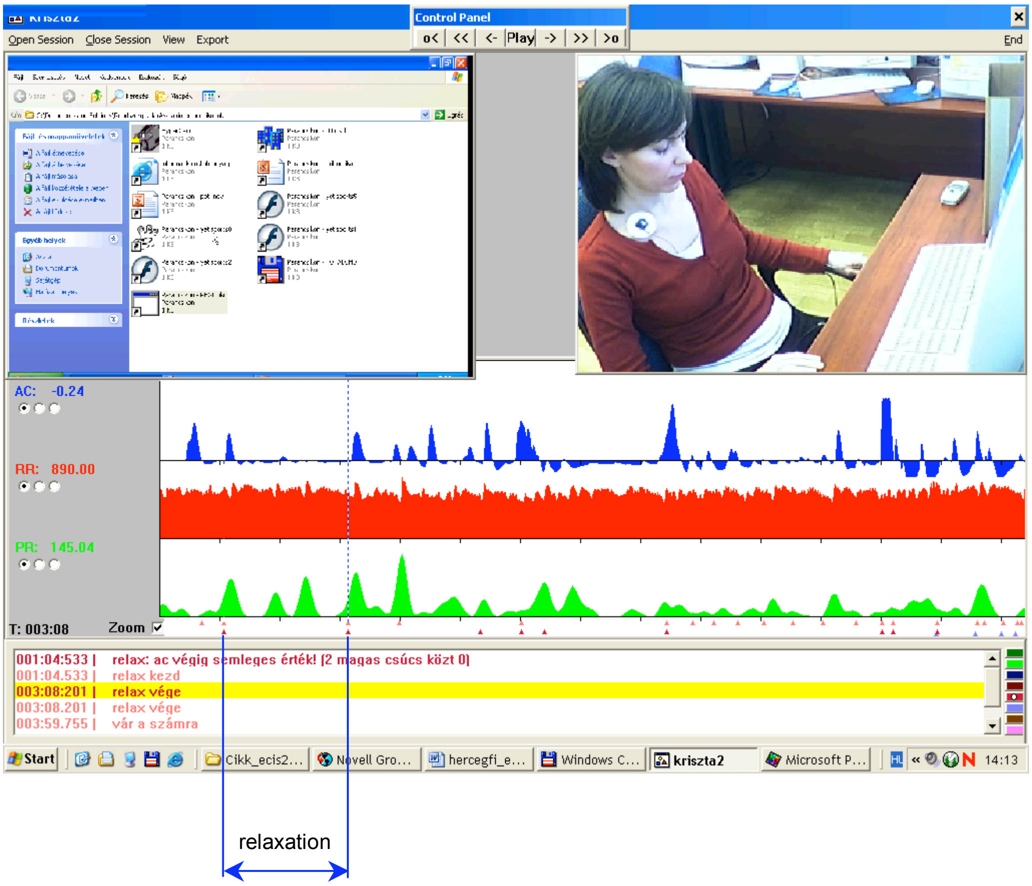

Figure 7. *The new version of INTERFACE frame system integrating a new physiological channel: Skin Conductance (SC).*
The upper blue curve shows the Alternating Current (AC) component of the SC.
The crosshair stands at the end of a period of relaxation.
In this case, the relaxation was not successful from the view point of mental effort (the lower green curve of HPV was fluctuating too much), but the emotional state of the subject was very calm: the upper blue curve of SC was really low and smooth.

5 SUMMARY OF EXPERIENCES GAINED THROUGH THE USE OF THE INTERFACE SYSTEM

The INTERFACE methodology is very useful if:

- the goal is to identify a great number of concrete – smaller or bigger – usability flaws in order to feed them back to the program designers and coders,
- the accurate knowledge of the interaction history (keystrokes, clicks, behaviour and screen content video recordings, physiological parameters, etc.) is important,
- the user has to perform complex tasks, while watching the video playback helps a lot to recall the problems and the episodes of cognitive load and related mental effort,
- the goal is to find design errors reliably.

Based on the results presented here as well as in related papers, it can be stated that the INTERFACE methodology in its present form is capable of identifying the relative weak points of the HCI. By this methodology and the related workstation, it was possible to study events occurring during the HCI in such detail and objectivity that would not have been possible using other methods presently known to us. The sophisticated Heart Period Variance (HPV) profile function integrated into the INTERFACE system is a powerful tool for monitoring events in such a *narrow time frame* that it can practically be considered as a time-continuous recording of relevant elementary events.

Acknowledgments

The authors would like to thank the students and teachers of the Richárd Kolos Technical High-School and the Elek Fényes Economical High-School for their participation, the developers of the multimedia material for their previous work, and Endre Levente Erdős, Zsuzsanna Hanzlik and Levente Csányi for their valuable contribution.

References

Chen, D., Vertegaal, R. (2004): Using Mental Load for Managing Interruptions in Physiologically Attentive User Interfaces. In Extended Abstracts of the CHI 2004 Conference on Human Factors in Computing Systems, 2004 April 24-29, Vienna, Austria. pp.1513-1516.

Izsó, L. (2000). Discrimination between Design Errors and User Errors by Binomial Test. Behaviour & Information Technology, Vol. 19, No. 5. pp. 379-384.

Izsó, L. (2001). Developing Evaluation Methodologies for Human-computer Interaction. Delft University Press. ISBN 90-407-2171-8, pp. 236, Delft, The Netherlands.

Izsó, L. and Hercegfi, K. (2004). HCI Group of the Department of Ergonomics and Psychology at the Budapest University of Technology and Economics. In Extended Abstracts of the CHI 2004 Conference on Human Factors in Computing Systems, 2004 April 24-29, Vienna, Austria, chapter of HCI Overviews.

Izsó, L. and Láng, E. (2000): Heart Period Variability as Mental Effort Monitor in Human Computer Interaction. Behaviour Information Technology, Vol. 19, No. 4. pp. 297-306.

Izsó, L., Mischinger, G. and Láng, E. (1999): Validating a new method for ergonomic evaluation of human-computer interfaces. Periodica Polytechnica, Social and Management Sciences. Vol. 7, No. 2, pp. 119-134.

Izsó, L. and Zijlstra, F. (1999). Efficiency in Work: An approach to interface evaluation and -design. WORC Report 99.10.001, ISBN 90-75001-24-X, New Approaches to Modern Problems in Work Psychology, pp 77-98, Work and Organization Research Centre, Tilburg, The Netherlands. (An extended version of a paper presented at the 8th European conference on Work and Organizational Psychology. Proceedings pp. 7-19. April 2-5, 1997, Verona, Italy).

Láng, E. and Szilágyi, N. (1991). Significance and assessment of autonomic indices in cardiovascular reactions. Acta Physiologica Hungarica. Vol. 78 (3) pp 241-260.
Luczak, H. and Laurig, W. (1973). Analysis of heart rate variability. Ergonomics, 16, 85-97.
Mulder, G. and Mulder-Hajonides van der Meulen, W. R. E. H. (1973). Mental load and the measurement of heart rate variability. Ergonomics, 16, 69-83.
Rowe, D. W., Sibert, J., and Irwin, D. (1998). Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. Proceedings of CHI98, 18-23.
Sayers, B. McA., (1971). The analysis of cardiac interbeat interval sequences and the effect of mental work load. Proceedings of the Royal Society for Medicine, 64: 707-710.