A multi-modal architecture for non-intrusive analysis of performance in the workplace

Davide Carneiro[a,b], André Pimenta[a], José Neves[a], Paulo Novais[a]

[a] *ALGORITMI/Department of Informatics, University of Minho, Braga, Portugal*
[b] *CIICESI, ESTGF, Polytechnic Institute of Porto, Portugal*

Abstract

Human performance, in all its different dimensions, is a very complex and interesting topic. In this paper we focus on performance in the workplace which, asides from complex is often controversial. While organizations and generally competitive working conditions push workers into increasing performance demands, this does not necessarily correlates positively to productivity. Moreover, existing performance monitoring approaches (electronic or not) are often dreaded by workers since they either threat their privacy or are based on productivity measures, with specific side effects. We present a new approach for the problem of performance monitoring that is not based on productivity measures but on the workers' movements while sitting and on the performance of their interaction with the machine. We show that these features correlate with mental fatigue and provide a distributed architecture for the non-intrusive and transparent collection of this data. The easiness in deploying this architecture, its non-intrusive nature, the potential advantages for better human resources management and the fact that it is not based on productivity measures will, in our belief, increase the willingness of both organizations and workers to implement this kind of performance management initiatives.

Keywords: Performance, Monitoring, Neural Network, Non-intrusive

1. Introduction

The change in the job offers in the last decades, caused by technological evolution, brings along many significant and broad changes. Some of the most notorious ones can be pointed out by the emergence of indicators such

as stress or mental fatigue which, in extreme cases, can endanger the life and well-being of the employees. In more moderate cases it will impair performance, general cognitive skills and productivity. In addition to these factors, many of these jobs are the so-called desk-jobs, in which people frequently sit for more than 8 hours [1].

Until now, the performance of the employees has been evaluated through their productivity: the more one produces, the better the performance at work. While the true nature of this relationship is yet to be thoroughly studied (properly contextualized in each work domain), there are other issues that need to be addressed. First of all, the worst aspect about this approach is that it only points out a potential decrease of performance after a productivity loss. This means that the "damage" is already done and that it is most likely too late for the employee to cope with whatever caused the performance loss and. An approach that could point out, in advance, upcoming breaks in performance (e.g. through the observation of behavioral patterns) could allow for preventive interventions rather than reactive [2].

Another major aspect that current approaches fail to consider are the side effects of productivity or performance monitoring in the workplace [3]. Indeed, as research in the last years has pointed out, this kind of approach might act as an additional stressor on the employee, which adds to the existing pressure in the workplace and to its negative consequences. In a study conducted in 1995 by researchers of the State University of New Jersey, it was analyzed the impact of electronic performance monitoring and its social context on the productivity and stress of employees [4].

Electronic Performance Monitoring (EPM) systems are one of the many technological developments employees face in today's workplaces. These systems provide managers a wide range of information about employees' routines including real-time information such as the pace of work, degree of accuracy, log-in and log-out times, and even the amount of time spent on bathroom breaks. This study examined how productivity and subjective experiences are affected by EPM systems and how the social context of the workplace moderates that influence.

In a survey involving the monitored workers, 81% of the respondents declared that electronic observation made their jobs more stressful [5]. Another study compared the behavior of monitored and non-monitored workers and found that monitored workers felt more stressful [6].

The introduction of EPM systems can transform ordinary jobs into high-stress jobs. It can also reduce the opportunities for employees to socialize

with each other at work, leading to a loss of social support, partially responsible for the stress associated with EPM [7, 8].

In this paper we present a new approach on the problem of performance monitoring in desk jobs, in line with the Ambient Assisted Living view[9], that quantifies performance independently of the amount of work produced by the employee. Namely, we develop a multimodal approach that incorporates different sources of information to allow the extraction of behavioral and physical features to characterize the performance of the user. These features are extracted from the keyboard, the mouse and the chair of the user.

As the results point out, the selected features vary consistently throughout the day, showing a decrease in the performance of interaction as the day of work goes by and an increase in the movement of the chair, pointing out increasing discomfort. This multimodal approach allows a quantification of performance decrements without requiring productivity measures. In that sense, employees will be more prone to accept such an approach. It is also non-intrusive as it requires no specific actions by the employees: they simply need to carry out their regular work.

Our goal is that team managers make use of such information to implement better human resources management strategies, that take into account (possibly in real-time) the state of the employees, allowing the development of individualized working schedules, warnings when performance decreases or the implementation of automatized coping strategies. As an example, we developed a simple desktop application that produces a warning when significant decreases in performance are observed on the user.

2. Architecture

The architecture of the proposed system was developed as a Service-Component Architecture (SCA): a group of OASIS specifications that has become an industry standard. It is intended for the development of applications based on SOA, which defines how computing entities interact to perform work for each other. Originally published in November 2005, SCA is based on the notion that all the functions in a system should exist in the form of services that are combined into composites to address specific business requirements. In other words, it allows to build service-oriented applications as networks of service components. SCA is used for building

service components, assemble components into applications, deploy to (distributed) runtime environments and reuse service components built from new or existing code using SOA principles.

SCA provides a good basis for applications under the umbrella of the Ambient Intelligence (AmI) field[10], such as this one, and it fulfills major AmI deployment requirements by promoting late bindings at deploy time and runtime with the support of several relevant technologies including POJO, SOAP, REST, BPMN, BPEL, JMS, Camel or Rules services. But most of all it is currently supported by several major commercial and open source products such as Jboss Switchyard, IBM WebSphere or TRENTINO (C++). From the several available implementations of SCA we have chosen JBoss SwitchYard since it is an open source solution in a relative mature state, and also enhances some of the SCA advantages.

A service-based approach was followed to develop an architecture logically divided into several packages that encapsulate a set of features and tasks. Figure 1 pictorially depicts, from a high-level point of view, the proposed architecture.

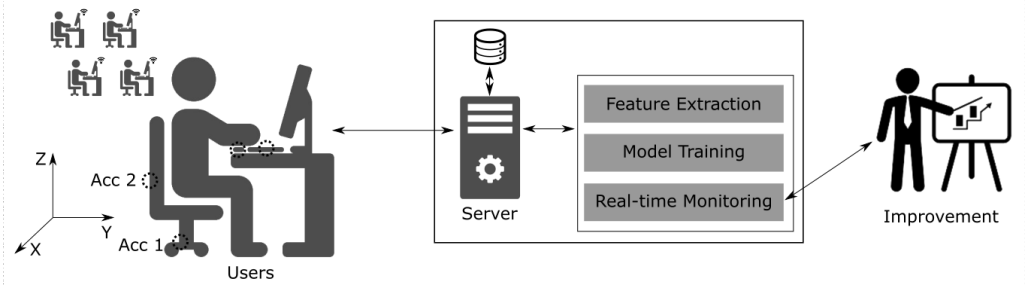

Figure 1: High-level view of the architecture, highlighting the 4 sources of information in each user: the mouse, the keyboard and two accelerometers placed on the chair as well as their placement and directions of the axes.

Three main components can be identified. The first is the component that is on the user-area. It provides, in an non-intrusive way, features about the employees' behavior. Two of these features are extracted from two accelerometers placed in the chair, as detailed in Figure 1. These accelerometers aimed to record the movements of the workers during the day (8 hour workday), while sitting in front of the computer. Accelerometer 1 was placed at the level of the worker's back while accelerometer 2 was placed in one of the wheeled arms of the chair, with the aim to record acceleration generated by

the moving of the chair. Specifically, Axivity's WAX3 wireless accelerometers were used.

The remaining features are extracted from the mouse and keyboard of the computer and fully characterize the employee's interaction with these peripherals. Although these features and the process of their extraction have been detailed in the past[11], we provide a brief overview. From the keyboard we extract features such as the typing speed or rhythm, the number of errors or the time of each key press. In previous studies we have noted that key presses, for example, may vary in duration from 80ms at the beginning of a session, to as much as 100ms at the end, depending on the user and on the type of task being performed. From the mouse we extract features such as velocity, acceleration, distance between clicks, duration of the click or excess of distance traveled, just to name a few. Once again, we consistently find decreases in performance on these features as the workday progresses [12].

Simultaneously with the acquisition of the behavioral features, employees answered a questionnaire about mental fatigue on a hourly basis (USAFSAM Fatigue Scale [13]). This was implemented with the aim of studying, in parallel, the daily evolution of mental fatigue given the well-known relationship between this indicator and performance [14].

The second component of the architecture is placed on the server side of the system. It is responsible for the continuous acquisition of behavioral data and its persistence in the database. Moreover, it also provides very important functionalities in the form of services. Namely, it allows for behavioral models to be trained based on individual user data and the results of the questionnaire (as depicted further below), namely using machine learning algorithms. It also allows for these models to be used in real-time for classifying user performance from behavioral data, which is essential given the aims of this research line.

Finally, the third major component of the architecture concerns the use of the performance models, in real-time, for improving performance throughout the day or preventing performance breaks. This considers both automatic or human-driven decision mechanisms. On the one hand, and as implemented in the developed prototype, the system can autonomously point out to the user or anticipate significant breaks in performance, providing a warning that encourages the user to take a break. Other more advanced context-dependent coping strategies can be implemented (e.g. taking a walk in a nearby park, grabbing a coffee, group activity with coworkers).

On the other hand, human decision makers can also make use of this

kind of information, which may or may not be directly shown to the employees, in order to better manage their teams. Namely, managers can learn the best working schedule for each individual or each employee's resilience to working long hours. This will allow them to implement better strategies in what concerns the management of personnel, ultimately achieving working environments that are better for both organizations (e.g. increased performance/productivity) and collaborators (e.g. no stress from productivity measures, better social relationships, improved quality of life).

Summarizing, in an organizational context, the gathering and analysis of metrics describing people's behavior, and the providing of tools for visualization (particularly real time analytics) enables better decision making and data-driven actions that consider the state and well-being of each individual worker. Such initiatives can nowadays be scaled to hundreds or thousands of workers, through the use of Big Data tools and techniques, without compromising performance and availability.

The interaction features described above characterize the behavior of each individual while interacting with the computer. This behavior, as many others, is affected by factors that influence performance at work, including mental fatigue, stress level or emotional arousal. Specifically, each instance of the behavior is characterized by fifteen values (represented as doubles) that are a result of applying several data summarization techniques (e.g. i.e. aggregation of collected data by calculating values such as mean and variance on the very frequently collected values). Each of these instances also contains a timestamp.

Given that this data is stored in a MongoDB databae, each record needs 136 bytes of storage space: 15 times 8 bytes (the MongoDB double size) plus 8 bytes for the timestamp, and 8 bytes for the two keys that describe the application being used and the user. A new record is produced every five minutes, for each user of the environment. Assuming that each individual is expected to work around 8h per day, a production of around 12.75 Kbytes of data per worker is estimated. Table 1 shows the expected data growth projections for different numbers of users and different time-spans.

3. Validation of the Architecture

The developed architecture was used in a laboratory setting with the main aim to assess a potential relationship between performance and the features under study, namely the acceleration measured on the chair and the

Table 1: Projections of data growth for different number of users and time frames.

Time \ # Users	1	100	10000	1000000
5 minutes	136 bytes	13.28 Kbs	1.297 Mbs	129.7 Mbs
1 day	12.75 Kbs	1.245 Mbs	124.5 Mbs	12.159 Gbs
1 week	89.25 Kbs	8.716 Mbs	871.6 Mbs	85.115
1 month	382.5 Kbs	37.354 Mbs	3.648 Gbs	364.8 Gbs
1 year	4.545 Mbs	454.5 Mbs	44.382 Gbs	4.438 Tbs

interaction with the peripherals. In what concerns the interaction with the peripherals, 24 participants (19 men) interacted with the computer while performing their regular tasks, with their interaction patterns being recorded. Of these, and due to hardware availability limitations, only 8 had accelerometers placed in their chairs, as depicted in Figure 1. As would happen in a real-life environment, these participants were requested to come into the lab and perform their regular activities, without any restriction whatsoever.

Data was collected continuously and organized on an hourly basis. Each hour of data originates one instance, with each instance depicting the variance of each feature in that period for a given user, as well as a subjective evaluation of each user's level of fatigue, provided through a self-report mechanism. Specifically, we used the seven-point USAFSAM Mental Fatigue Scale created by Dr. William F. Storm and Captain (Dr.) Layne P. Perelli of the Crew Performance Branch of the USAF School of Aerospace Medicine, Brooks AFB, San Antonio, Texas, and then used in many field and laboratory tests [15].

Data from the accelerometers was organized in hourly series, so as to match the intervals at which questionnaires were collected. Using the Anderson-Darling test, it was determined that the distributions of the eight resulting datasets were not normal. Given this, the Kruskal Wallis test was used in the subsequent analysis. This test was used to compare the distributions of each of the three axes and of the two accelerometers. The maximum p-value observed was 2.2e-16, which demonstrates that the changes in acceleration that occur from one interval to the next are statistically significant.

The difference in the distributions can also be visually observed as in Figure 2. This figure depicts violin plots, which in addition to the information provided by boxplots include the probability density for the different values.

These plots show the evolution of the distributions of the values during the day.

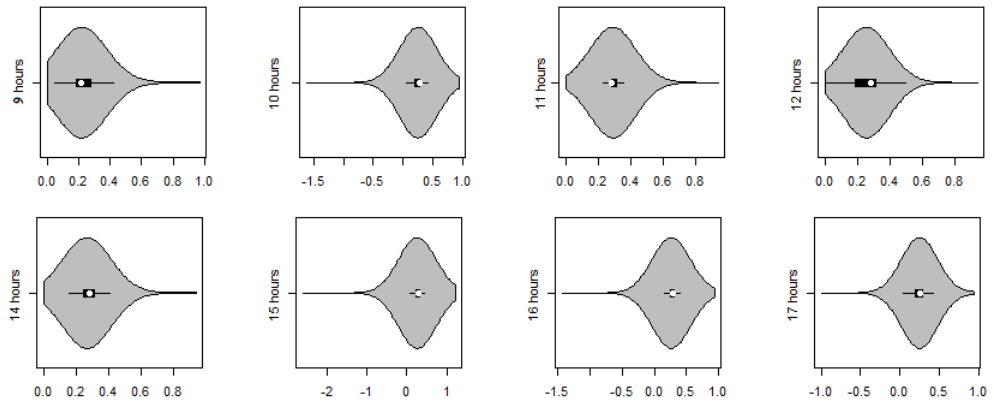

Figure 2: Hourly Violin Plot of x-axis for accelerometer 2.

Having carried out this preliminary analysis of the data, it was examined whether there is any relationship between the subjective feeling of mental fatigue and the acceleration measured on the chair, in both positions. To this end it was used the Pearson's test to determine the statistical correlation between the acceleration measured in each axis and the subjective level of fatigue.

Data show that the readings of accelerometer 1 have a positive correlation with the level of mental fatigue for the axes X and Y (0.67 and 0.53, respectively). Concerning accelerometer 2, a strong negative correlation with the subjective feeling of fatigue exists for the y-axis (-0.9).

The Mann-Whitney test was also used to compare the distributions of the values of each feature extracted from the mouse and the keyboard, in order to verify that there were statistically significant difference between the two groups of data (Table 2).

4. Results

User feedback was paired with the variance of the features considered in each hourly period to train the neural network and thus understand the relationship between the variables. In the preliminary analysis of the data of the accelerometers, a relationship between the subjective levels of mental fatigue

Table 2: Analysis of significance of the collected data. Column a depicts the average value of the Mann Whitney test. Column b depicts the percentage of individuals for which the differences observed in each feature when fatigued and rested were statistically significant.

Features	a	b
Distance of the Mouse to the Straight Line	0.022	79%
Key Down Time	0.021	75%
Mouse acceleration	0.050	67%
Average Distance of the Mouse to the Straight Line	0.042	63%
Mouse velocity	0.023	63%
Average Excess of Distance	0.041	54%
Time Between Clicks	0.020	54%
Distance During Clicks	0.026	54%
Time between keys	0.271	46%
Total Excess of Distance	0.258	46%
Double Click Duration	0.003	33%
Absolute Sum of Angle	0.336	25%
Signed Sum of Angles	0.706	21%
Distance between clicks	0.0238	20%

and the movement of the chair in some of its axes in both accelerometers was found. For modelling accelerometer 1, which showed a strong negative correlation with fatigue for axes X and Y and no significant correlation for the z-axis, a linear regression was fit to the average acceleration values on these axes. While the z-axis is left out due to a lack of correlation, its absence is also positive in the sense that acceleration from when the user sits down or stands up (which happens mostly on this axis and is not the normal behaviour we are observing) will not influence the model.

The trained linear model contains residual values shown in Figure 3 where it is possible to verify that these values follow a normal distribution, and they have a median value close to 0. In this model we can still observe a Multiple R-squared: 0.8385, Ajusted R-squared: 0.8331 which confirms the correlation shown between the acceleration in the y-axis and the level of fatigue.

For the data recorded by the accelerometer 2, K-means clustering was used to verify that the acceleration of the chair movement in all three axes can be associated with the different levels of fatigue observed. The value of K was selected using the NbClust package from the R Software [16]. This package

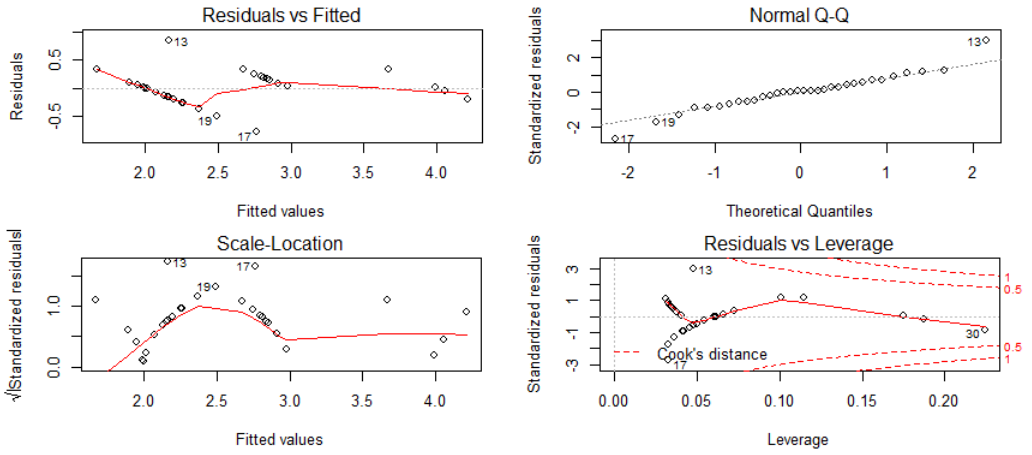

Figure 3: Linear model built for accelerometer 2.

provides 30 indexes for determining the number of clusters and proposes the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. The suggested value of K with the largest number of indexes chosen was three. Therefore a 3-cluster solution was proposed.

The accuracy of the resulting clusters was computed through the adjusted Rand index. The adjusted Rand index provides a measure of the agreement between two partitions, adjusted for chance. It ranges from -1 (no agreement) to 1 (perfect agreement). Agreement between the labeled data and the cluster solution is 0.56.

The results obtained for accelerometer 2 show that it is possible to separate the collected data into three distinct clusters. While workers pointed out four different levels of fatigue, the first level was only chosen 10% of the times. This division thus make sense also when considering the answers to the questionnaires.

In what concerns the features extracted from the mouse and keyboard, the first step in the process of training a suitable classifier was to select the features that presented the most significant differences when comparing the two groups of data. From all the available features, we selected the following: key down time, mouse velocity and acceleration, time between clicks, distance from the pointer to the line between clicks, distance during clicks, average excess of distance and average distance of the mouse to the straight line and the time of the day (morning, afternoon, evening) to rule out the influence

of the circadian rhythm.

As a classification algorithm we selected an Artificial Neural Network (ANN). This decision was taken after an analysis of existing work that showed the suitability of this kind of algorithms for modeling mental fatigue. Namely, existing works use electroencephalography signals [17], electromyography [18] and other features such as gaze detection or face pose [19] as inputs to neural networks. Existing related work is especially targeted at vehicle driving and operation of machines, which is a classical field of application of fatigue detection and performance monitoring [19]. To the extent of our knowledge, it is the first time that an ANN is used with performance features such as those put forward in this work. This is fundamental for domains such as the workplace or classrooms, in which the use of invasive approaches results impractical.

The previously mentioned group of features thus constitutes the input layer of the neural network. Its output is a value between 1 and 7 that denotes the degree of the level of fatigue, as in the USAFAM questionnaire.

With the input and output layers defined, a multilayer feed-forward neural network was used to model the relationship between behavioural features and a level of fatigue. The R software environment was used. A back-propagation learning algorithm was used which, besides the input and output layers, uses a intermediate hidden layer that lies between the input layer and the output layer, with a total of 10 nodes. The number of hidden nodes has been selected after the carrying out of several performance tests with different numbers of hidden nodes. In what concerns activation and error functions, the algorithm's default settings were used.

The network was trained with a dataset containing a total of 74 instances, each of which containing the value of the variance of each of the features over a period of one hour and the level of subjective fatigue provided by the users through the questionnaire, for that period. The network was trained during 125 iterations where it reached a minimum RMSE (Root-Mean-Square Error) as can be seen in Figure 4.

The ANN trained and outlined above was tested and validated with data from the second week of the data collection period, in which each instance, as described above, contained the variance of each feature in the period of an hour and the subjective measure of the level of fatigue of the user. This approach allows us to compare the value pointed out by the user against the value provided as output by the ANN. The main result is that the trained ANN correctly classified 81% of the instances, i.e., producing as output the

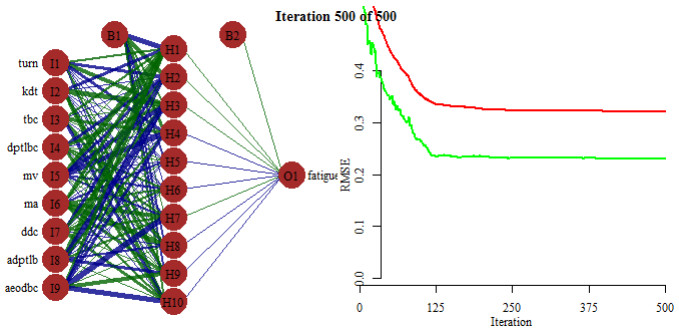

Figure 4: Evolution of the RMSE during the evolution of the training of the ANN (green line) and during validation (red line).

same value that the user provided in the questionnaire. The 19% that were misclassified where nonetheless classified as neighbouring values

5. Conclusions

This paper presented a distributed architecture for the non-intrusive acquisition of interaction and behavioral features for the classification of human performance. We have shown that certain axes of accelerometers placed on specific points of a chair as well as some of the interaction features extracted from the mouse and keyboard do correlate with performance. The most positive aspect of this work is that it allows the classification of human performance without the use of features that are not positively regarded by workers, namely productivity measures.

In that sense, workers will be more prone to accept this kind of performance monitoring. From the point of view of the organization, it is a inexpensive approach and does not require any change to the workers' routines, which is positive. Moreover, it will provide valuable information to team managers, namely concerning each of the workers' natural rhythms (e.g. some work better in the morning, others in the afternoon) thus allowing for a more effective management of human resources.

In the overall, this will result in more sensitive and positive working environments, with expected positive impacts in productivity measures as well as the quality of the product and of the workplace.

The ANN developed in this work is a part of a wider performance model being developed by this research team which includes related aspects such as

the performance of interaction and the degree of attentiveness. Ultimately, we aim at the development of a theoretical and practical multi-modal model that details how performance in the workplace is influenced by external factors and how it can be assessed non-intrusively.

Acknowledgments

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. The work of Davide Carneiro is supported by a Post-Doctoral Grant by FCT (SFRH/BPD/109070/2015).

Bibliography

[1] M.-H. Liao, C. Drury, Posture, discomfort and performance in a vdt task, Ergonomics 43 (3) (2000) 345–359.

[2] D. Carneiro, P. Novais, F. Andrade, J. Zeleznikow, J. Neves, Using case-based reasoning and principled negotiation to provide decision support for dispute resolution, Knowledge and information systems 36 (3) (2013) 789–826.

[3] L. Lima, P. Novais, R. Costa, J. B. Cruz, J. Neves, Group decision making and quality-of-information in e-health systems, Logic Journal of IGPL (2010) 315–332.

[4] J. R. Aiello, K. J. Kolb, Electronic performance monitoring and social context: impact on productivity and stress., Journal of Applied Psychology 80 (3) (1995) 339.

[5] L. Gallatin, Electronic Monitoring in the Workplace: Supervision Or Surveillance?, CNOT (241 St. Botolph St., Boston, MA 02115), 1989.

[6] R. Irving, C. A. Higgins, F. R. Safayeni, Computerized performance monitoring systems: Use and abuse, Communications of the ACM 29 (8) (1986) 794–801.

[7] M. J. Smith, P. Carayon, K. J. Sanders, S.-Y. Lim, D. LeGrande, Employee stress and health complaints in jobs with and without electronic performance monitoring, Applied Ergonomics 23 (1) (1992) 17–27.

[8] B. C. Amick, M. J. Smith, Stress, computer-based work monitoring and measurement systems: A conceptual overview, Applied Ergonomics 23 (1) (1992) 6–16.

[9] D. Carneiro, R. Costa, P. Novais, J. Neves, J. Machado, J. Neves, Simulating and monitoring ambient assisted living, in: Proceedings of the ESM, 2008, pp. 175–182.

[10] P. Giner, V. Pelechano, An architecture to automate ambient business system development, in: Ambient Intelligence, Springer, 2008, pp. 240–257.

[11] D. Carneiro, P. Novais, J. M. Pêgo, N. Sousa, J. Neves, Using mouse dynamics to assess stress during online exams, in: Hybrid Artificial Intelligent Systems, Springer, 2015, pp. 345–356.

[12] A. Pimenta, D. Carneiro, P. Novais, J. Neves, Detection of distraction and fatigue in groups through the analysis of interaction patterns with computers, in: Intelligent Distributed Computing VIII, Springer, 2015, pp. 29–39.

[13] L. P. Perelli, Fatigue stressors in simulated long-duration flight. effects on performance, information processing, subjective fatigue, and physiological cost, Tech. rep., DTIC Document (1980).

[14] R. Kanfer, Determinants and consequences of subjective cognitive fatigue., in: Cognitive fatigue: Multidisciplinary perspectives on current research and future applications, American Psychological Association, 2011, pp. 189–207.

[15] J. C. Miller, Cognitive performance research at brooks air force base, texas, 1960-2009.

[16] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: An R package for determining the relevant number of clusters in a data set, Journal of Statistical Software 61 (6) (2014) 1–36.

[17] L. King, H. T. Nguyen, S. Lal, Early driver fatigue detection from electroencephalography signals using artificial neural networks, in: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, IEEE, 2006, pp. 2187–2190.

[18] A. Subasi, M. K. Kiymik, Muscle fatigue detection in emg using time–frequency methods, ica and neural networks, Journal of medical systems 34 (4) (2010) 777–785.

[19] Q. Ji, X. Yang, Real-time eye, gaze, and face pose tracking for monitoring driver vigilance, Real-Time Imaging 8 (5) (2002) 357–377.