Cognitive workload and affective state: a computational study using Bayesian networks

P. Besson*, C. Maïano†, M. Durand*, L. Bringoux*, T. Marqueste*,
D. R. Mestre* C. Bourdin*, E. Dousset * and J.-L. Vercher *
*CNRS & Aix Marseille Université
UMR 6233 Institute of Movement Sciences, 13288 Marseille, France
Email: patricia.besson@univmed.fr
†Cyberpsychology Laboratory
Department of Psychoeducation and Psychology, University of Quebec in Outaouais (UQO), Canada

***Abstract*—This paper uses Bayesian networks to investigate the impact of three different kind of inputs, namely, physiological, cognitive and affect features, on workload estimation, from a computational point of view. The ability of the proposed models to infer the workload variation of subjects involved in successive tasks demanding different levels of cognitive resources is discussed, in term of two criteria to be jointly optimized: the diversity, i.e. the ability of the model to perform on different subjects, and the accuracy, i.e., how close from the (subjectively estimated) workload level the model prediction is.**

I. Introduction

Operators involved in complex multitask activities, such as piloting a helicopter, must constantly make quick and relevant decisions. Advanced systems provide them with some assistance, by delivering information on the task's context and by automating some processes. However, these automated agents also impose new information processing demands and might thus increase the level of cognitive workload (denoted as *workload* from now on)[1]. Therefore, intelligent systems, able to adapt to the current level of operators' workload might be more efficient. Such systems should provide greater assistance in case of overload, but should delegate more functions to the operator in case of low workload (likely to result in a lack of vigilance) [2], [3], [4].

Computational models have been proposed to infer cognitive states, such as workload or distraction, from task performance analyses or sensorimotor features (gaze, head movements, etc.) [5], [6], [7]. However, for these features to make sense, they have to be compared to nominal values that are dependent on the task context. More direct and task independent features can be extracted from physiological measurements. Indeed, changes in the subject's cognitive state may result in changes in physiological data [8], specifically (but not exclusively) when they are under the control of the autonomic nervous system (ANS). The latter is responsible for maintaining the body's homeostasis, noticeably through the orthosympathetic branch which mobilizes cardiorespiratoy and energy resources in response to the changing demands of the external and internal milieu [9]. Thus, electrocardiogram (ECG), electromyogram (EMG), skin conductivity (SC), and respiration were used in [10] to infer the stress level by drivers using linear discriminant analysis. In [11], the authors developed an Artificial Neural Network (ANN) taking electroencephalogram (EEG), electrooculographic (EOG) and respiration as inputs to assess workload levels. ECG, EEG and EOG were also used in [12] to derive an information-theoretic indicator of cognitive state. Support Vector Machine and ANN were applied to workload estimation in [13], using EEG, SC, respiration, and heart rate (HR) data. These models rely on physiological measurements to infer cognitive states but not necessary workload specifically (for example, stress, inferred in [10], should not be confused with workload thought it partly results from overload [14]).

Additionnally, several works, taking place in the field of affective computing, have addressed the problem of inferring users' affects from physiological signals [15], [16]. Indeed, it has been shown that humans' emotional intelligence provides them *the capacity to reason about emotions, and of emotions to enhance thinking* [17]. Thus, Isen et al. [18] showed that positive affect enhance problem solving and decision making. Therefore, affective computing aims at endowing machines with emotional skills, in particular, the ability of perceiving and adapting to user's current affective state to improve the efficiency of human-machine interfaces.

The objective of this work was to use Bayesian networks (BN) to study the contribution of physiological, cognitive and affective features on workload estimation, from a computational point of view. We firstly collected representative data through a dedicated experimental protocol. Since we wanted to use non-invasive and minimally intrusive sensors, we restricted the measurements to EMG, HR, SC and respiration and did not measure EEG (incompatible with helmets wore by helicopter or fighter pilots for example). Entropy values of these signals define the physiological inputs of our models. The reaction time (RT) to a secondary task is used as a cognitive feature [19], [3]. Finally, the positive and negative affect scale (PANAS) [20], [21] was used to evaluate the participants' subjective affective state during the experiment.

Different BN structures, built from expert knowledge, are tested, using in turn several combinations of these features. Their performances are evaluated in term of two criteria to be jointly optimized: the diversity (i.e., the ability of the model to be functional for different subjects), and the accuracy (i.e., how

close from the workload level the model prediction is). The ground truth is provided by subjective evaluations of workload collected during the experiment.

Rather than assessing the ability of the BN models to infer the *workload level*, we focused on how good they are in predicting the *workload variation* between successive tasks. Indeed, in the context of defining adaptive intelligent systems, an erroneous prediction of the workload level resulting in a false prediction of workload change (i.e., in a predicted variation opposite to reality) should be absolutely avoided. It might drive the system to undertake actions opposite to those required by the operator's state, with dramatic consequences.

Section II describes the experimental protocol used to collect representative data. An analysis of these data is also performed in this section to ensure that the subjects' workload has been manipulated by the experimental paradigm. The proposed models are presented in sec. III and their performances are assessed in sec. IV in term of two criteria to be jointly optimized: the diversity and the accuracy.

II. Material and Method

A. Subjects

Ten subjects (9 males and 1 female, aged 30 ± 10.7 years) with normal or corrected to normal hearing and seeing, participated in the experiment.

B. Material

The subjects sat in the dark, facing a standard 24" monitor, where graphical dynamic flying scenes generated by the home-grown ICE software [22] were displayed. An experimenter's computer was used to acquire all the data synchronously, using the Captiv Software [23]. These data were made of the simulation data (e.g., aircraft position) sampled at 100Hz, and of physiological data, acquired at a sampling rate of 2048Hz using the FlexComp Infinity sensors and encoder [24]. The subjects bore stereo headphones, so that they could hear pre-recorded instructions (the instructions' tone and content were then strictly identical for each subject) and the task related noises such as the engine noise (leading to a greater immersion) or the possible alarms.

C. Procedure

Using a regular joystick, subjects were asked to pilot a flying aircraft and to do their best to follow a trajectory defined by 60 rings, alternatively red and yellow. The trajectories varied only along the vertical dimension. The aircraft's speed was maintained constant at the same predefined value for all the trajectories. The ratio of hit rings over the total number of rings in the trajectory appeared on the cockpit dashboard. There was also a green or red light indicating whether the last ring had been hit or missed.

The experiment was organized in 5 sessions of 6 trials. Each trial lasted approximately 90sec[1]. In the first three sessions (labeled *D1A0*, *D2A0* and *D3A0*), the subjects were presented with three different trajectories of increasing difficulty (D1, D2, and D3), that remained the same for the 6 trials of each session. The trajectory difficulty was an independent variable meant to manipulate the task workload requirement. It was varied by changing the vertical distance between two successive rings, while keeping their depth distance constant. In the last two sessions (labeled *D1A1* and *D3A1*) the subjects were asked to fly again on the simplest and the hardest D1 and D3 trajectories, and to try to beat their own mean scores over these trajectories. Moreover, a strident alarm was emitted in case of a missed ring. This challenge and the alarm were introduced in order to maintain the subjects' motivation and implication in the task.

For each of the five sessions, a secondary task was introduced. Two geometrical shapes (a square or a triangle) appeared on the screen during 1sec, at pseudo-random positions (the ring apparition zone was avoided, and the same number of targets appeared in each of the four screen quarters) and at pseudo-random times (no apparition while the ring was crossed, and minimum time interval of 1.5sec between two successive targets). The subjects had to press a button on the joystick with the forefinger as quickly as possible in response to the square target apparition. They should not react to a triangle target.

Fig. 1 shows a typical screen shot of the simulated scene.

Fig. 1. Screen shot of a typical flying scene created by ICE. The ratio of hit rings over the total number of rings in the trajectory appeared on the cockpit dashboard (e.g. 1/60) and a green or red light indicated whether the last ring had been hit or missed.

D. Dependent variables

Performance on the primary (percentage of hit rings) and on the secondary tasks (false and good detection rates; RT) were recorded. The physiological variables comprised the following measurements:

- HR, estimated from the ECG by the Captiv software, using R-R intervals;
- Root mean squares of the flexor digitorum EMG (RMS1) and of the right trapezius descendens EMG (RMS2);
- Respiration (R), measured through chest expansion;
- SC, measured using electrodes placed on the first and little fingers of the left hand (temperature in the room equal to $19.33 \pm 0.98°C$).

Finally, psychological data were also collected at the end of each session. The subjects self-assessed their own workload during the performed task, using the NASA Task Load Index

[1]Though the speed is maintained constant, the duration of each trial is not necessary the same, since the aircraft's trajectory can be more or less sinusoidal

(TLX) scale [25]. The NASA TLX asks the subjects to rate their perceived workload on six different subscales. At the end of the experiment, these six components are matched two by two and the subjects have to choose for each couple which component best described the workload in the performed task. Each component score can thus be weighted accordingly to the number of times it has been chosen in the matching phase. In the present experiment, the NASA TLX rates on the six subscales are weighted and summed for each sessions to result in a single *Workload Index (WI)* (normalized on $[0, 1]$) per session. The subjects also filled the PANAS [20], [21] before the experimentation and after each session. This instrument was used to provide information on participants' affective reactions to the experiment. It comprised two affects' scale: positive (interested, excited, etc.), and negative (distressed, scared, etc.). The 20 items of this instrument were rated on a five point Likert scale from: (1) = "very slightly or not at all" to (5) = "Extremely". The ratings for each participant were normalized between 0 and 1, and the positive affect scores (PA) reversed ($1 - PA$) for all the scores to be interpretable in a consistent way. These normalized negative and reversed positive scores were then summed and normalized on $[0, 1]$, resulting in the *Affect Index (AI)* used in the following of the paper.

E. Analysis of experimental data

The impact of the primary task difficulty on performance in both the primary and secondary tasks is assessed through analysis of variance (ANOVA) statistical tests. There is no significant difference between trials for a same level of difficulty (p=0.13) whereas the difference is significant between sessions (F(4,36)=26.708, p=0.000). A post-hoc analysis (Student Neuwman-Keuls) indicates that scores are statistically different for the different difficulty levels ($p < 0.001$). The difficulty level also impacts the RT on the secondary task (F(4,24)=14.084; p=0.0000) whereas there is no intra-session effect (p=0.03). Finally, the TLX scores also increase with difficulty's session: differences are significant between sessions (F(4,26)=12.284, p=0.000). This statistical analysis establishes that subjects experienced different levels of workload during the task. This should be reflected in the physiological data and be captured by the model.

The analysis of the AI shows that variations of affects during the experiments are quite different from one subject to the others. Generally speaking, the variations are small: the maximal variation over all sessions is observed for subject 5 and is equal to 15%. Some subjects (noticeably, the subject 5) expressed negative affects prior to the task. For some subjects, the negative affects tend to decrease with the task difficulty, whereas it is the opposite for others. Finally, this analysis do not point out any striking consistency between affect variations and difficuly or workload (ass assessed by WI) variations (see e.g., subject 4, on Fig. 2). Therefore, it is not sure that adding knowledge about the subjects' affects to the model might improve its ability to infer workload.

III. Model

A. Selection of output and input features

A data-driven approach to modeling problem requires first to assign some data with the correct class labels so that the relationship between the input features (derived from the physiological data in the present case) and the classes (the model's output) can be automatically discovered and extracted in the learning phase. As stated in sec. II-E, we can only rely on a subjective rating scale to label our ground truth. This adds some noise in the pattern recognition process. To reduce the noise in the process as much as possible, the input features have also to be optimized: the more representative the features, the simpler the task for the classifier, thus the better its expected performance [26].

We expect (and we visually observed) changes in the variability or stability of the physiological signals with changes in the subjects' cognitive states. Indeed, the physiological signals we recorded are mainly under the control of the ANS, which regulates the body's homeostasis through successive activations of the sympathetic and parasympathetic systems (resulting in mobilization or slowing down of the organism) [27]. These changes can be captured by the Shannon's entropy of the physiological data, which will be the input for our model. To the best of our knowledge, such features have never been used in this context. The entropy of the random variable (rv) X is a measure of the average uncertainty in X [28]. Stated in a different and simpler way, it is a measure of disorder. Before the entropies to be estimated, the noise in the raw signals is firstly smoothed using a low-pass median filter. The first and last seconds of each trial's signal are also removed to avoid possible starting and ending effects. Then, the data are normalized between 0 and 1, taking the minimal and maximal values observed on the first three sessions (used as training sessions). Entropies are estimated on 15sec long windows slided by 5sec along the signals, using an histogram of 41 bins that ranges on $[0, 1]$. Therefore, there is about 90 values per sessions and per subject. Entropy values are also normalized between 0 and 1 by taking the maximal and minimal values over the first three sessions for each subject.

It can be observed that the variation of the mean entropy features is consistent with the variation of the performance on the primary task, the RT and the WI (see Fig. 2 showing the subject 4's features as an illustrative example). However, we observed the variation of the physiological data to be idiosyncratic: e.g., the mean entropy values of SC might decrease for some subjects or increse for some others (as for the subject 4) with the difficulty level. As a results, the models will be individual (trained and tested on each subject separately).

B. Model definition

BN are defined to infer the subject's WI on each session, from different kind of features. Firstly, we examine the impact of the number and type of the physiological inputs on the model performance. To this end, different classifiers are tested,

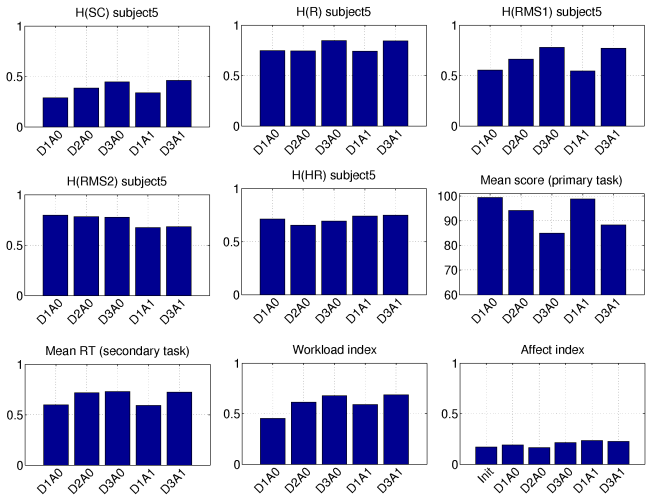

Fig. 2. Mean physiological feature values (entropy H of the signals), performance on the primary task, reaction times (RT), workload and affect index for each session performed by subject 4.

each taking one, two or three of the possible physiological features SC, R, HR, RMS1 and RMS2[2] as inputs. As a result, 25 classifiers with different physiological nodes as direct parents of the WI node (structure 1) are trained and tested.

Then, the impact of adding to the model a cognitive feature, namely, the RT, is evaluated. Whereas in BN with structure 1, WI node was a direct child of the physiological nodes, it is now a child of the RT node, which is itself a child of the physiological nodes. This defines the structure 2.

Finally, we wanted to assess how the affective state of the subject could impact the physiological data, and so, the prediction of the WI score. Therefore, the AI is also introduced as a possible parent of the physiological nodes in both structures. The different model structures are presented on Fig. 3 for a classifier made of two physiological nodes Φ_1 and Φ_2.

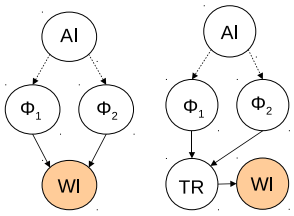

Fig. 3. Bayesian Network models inferring the WI value from two physiological features Φ_1 and Φ_2 either directly (Structure 1) or via RT (Structure 2). The models are tested with or without a AI node as parent of the Φ nodes (thus, this link is dashed on the graph). There can also be either a single or three physiological nodes.

The joint probability density functions (pdf) described by the BN are estimated on the training set using histograms with the following parameters (rv take on values in $[0, 1]$, but RT, taking on values in $[0, +\infty[$): 5 bins of width 0.2 for the physiological rv, 20 bins of width 0.05 for WI and AI, and 16 bins of width $\exp^{(0.2)}$, with the first bin centered on $\exp^{(-3.7)}$ and the last bin taking all the values greater than $\exp^{(-0.9)}$ for RT. For each subject, the training set is made of the data collected on the first three sessions *D1A0*, *D2A0* and *D3A0*. The testing set is made of the two last sessions *D1A1* and *D3A1*. Both the learning and inference stages have been implemented using the Bayes Net Toolbox for Matlab [29]. Because there are some missing data (HR in particular could not be reliably recorded sometimes, and there is not necessary one RT value per measurement window), the Expectation Maximization algorithm has been used with a stopping criterion of 10 iterations).

[2]For simplification purposes, the rv denoting the entropy features are named as the acronyms of the corresponding physiological data

C. Assessing the model performances

The performance of the models is assessed by looking at the differences in the WI between *D1A1* and *D3A1* sessions. The model output will be deemed as correct if the observed and inferred WI are evolving the same way, that is, if the performance index ρ, defined as follow, is positive:

$$\rho = \text{sign}(\Delta) \cdot \text{sign}(\Delta^*) \cdot |\frac{\Delta^*}{\Delta}|, \quad \text{if} \quad \Delta^* < \Delta \quad (1)$$

$$= \text{sign}(\Delta) \cdot \text{sign}(\Delta^*) \cdot |\frac{\Delta}{\Delta^*}|, \quad \text{else}, \quad (2)$$

where Δ is the difference between the subjects' WI on sessions *D3A1* and *D3A1*, and Δ^* the difference between the predicted WI on these two sessions. The quality of the model performance is given by the distance to 1 (the closer, the better). A fine analysis of the model's false detections is useless since we want this false detection rate to be null. Indeed, as stated in sec. I, a false estimation of the workload variation between successive tasks can not be accepted: it would lead the assistance system to undertake unadapted measures, which could have worse consequences than doing nothing.

For each model, we are looking at the performance over the subject set. Thus, we want the maximum number of subjects to be correctly detected, with a ρ score as close as possible to 1. This is a two-variable optimization problem, where the first parameter to be optimized is the model diversity and the second, its accuracy. The model diversity is assessed by looking at the percentage of subjects correctly detected (S). Also, to allow for comparisons between the accuracy performances, θ, the normalized area under the ρ curve, plotted as a decreasing function of S, is used rather than ρ:

$$\theta = 10 \cdot \frac{\sum \rho}{S}, \quad \theta \in [0, 1]. \quad (3)$$

IV. Results

From Fig. 4, it can be seen that most of the models are performant in either one of the two *diversity* (S) or *accuracy* (θ) criteria. However, we are interested in the models performant in both dimensions simultaneously. Depending on the criterion that is retained, the best model will be different. Table I presents the models that perform best on both the accuracy and diversity criteria, using a trade-off favouring alternatively each criterion. The trade-off is obtained by looking at the best

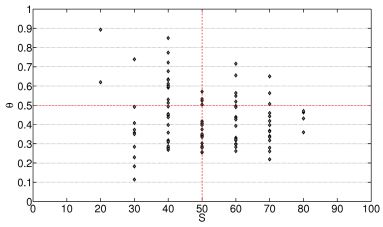

Fig. 4. Performance of the models in terms of diversity (S score) and accuracy (θ score). The best models lie in the upper right-hand side quarter of the graph.

TABLE I
BEST MODELS IN TERM OF A TRADE-OFF PERFORMANCE BETWEEN DIVERSITY AND ACCURACY, FAVOURING EACH OF THESE TWO CRITERIA ALTERNATIVELY

Best θ score with a S score of at least 50%					
Performance criterion	Structure	Physiological nodes	S	θ	AI node
Maximal Accuracy θ	1	SC;R;RMS1	60	0.72	Yes
Best S score with a θ score of at least 50%					
Performance criterion	Structure	Physiological nodes	S	θ	AI node
Maximal Diversity S	2	HR;RMS2;RMS1	70	0.65	Yes

model in term of accuracy (respectively, diversity) in the subset of models that obtain a score of at least 50% on the diversity (respectively, accuracy) criterion.

To analyze the impact of the number of physiological inputs, of RT and of AI on the models' performance, we will look at the set of best models, that is, the models with a performance greater than 50% for both diversity and accuracy criteria (Figs. 5 and 6). Notice that these are the models lying in the upper right hand side quarter of Fig. 4.

It is obvious from Figs. 5 and 6 that adding the AI node (i.e. information about the subjects' affective states) improves performance the most. Thus, when there is a AI node, 20% of the models fit the "good model" criterion with two physiological nodes only, and this number reaches 30% in case of three physiological nodes. Without the AI node, three physiological nodes are required to get some good models (10% only) and observation of the Fig. 6 shows that these models also contain the RT node (structure 2) since there are no good model with structure 1 and without AI. When structure 2 is used (whatever the number of physiological nodes) and a AI node included, the percentage of good models reaches 32%.

V. DISCUSSION AND CONCLUSION

In this paper, BN are proposed to infer the variation of workload for operators involved in multitask activities, from physiological, cognitive, and affective features. The primary objective is to examine, from a computational point of view, the impact of these different types of information on workload estimation. Therefore, two BN structures with different number of inputs are tested and compared in term of two criteria

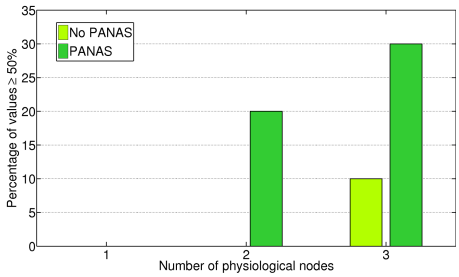

Fig. 5. Percentage of good models (with performance greater than 50% for both accuracy and diversity criteria) for the different physiological node numbers (no distinction is made between the two possible BN structures). Models with or without the AI node are compared.

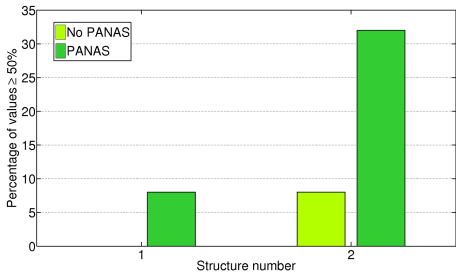

Fig. 6. Percentage of good models (with performance greater than 50% for both accuracy and diversity criteria) for the different structures (whatever the number of physiological node). Models with or without the AI node are compared.

to be jointly optimized: the accuracy and the diversity of the classifier.

The most striking result of this study is to put forward how taking into account subjects' affective state increases performance on workload estimation. Indeed, increasing the number of physiological inputs, and adding cognitive information (RT) in the model increase the number of "good models" (models with a performance greater than 50% on both the accuracy and diversity criteria), as could be expected from literature (see e.g. [30], [31]). But coupling these conditions with some knowledge about the subjects' affects (through the use of features derived from the PANAS questionnaire outputs) outstandingly improves models' performance.

Since relationships between differences in affective states and physiological changes have been shown in some studies (seee e.g., [32]), we suggest that providing the model with information about these affective states help it in getting rid of the physiological variations unrelated to subjects' workload changes. Of course, in this exploratory work, we are asking the model to estimate the subjects' state at the end of a task, giving it some information collected after this task has been performed. Therefore, as it is, the model cannot be used as a workload predictor. The purpose of this study was rather to point out that, since the relationships between psychological and physiological data is not one to one [8], it is essential to design the modeling process so that the physiological changes related to workload can be dissociated from the changes related to other psychological states.

The analysis of the performance of the different models points out that each of the five proposed physiological features might appear in classifiers with good performance. In particular, the two best models (obtaining the best performances on the two criteria θ and S, with a trade-off favoring ether one criterion or the other) do not rely on the same physiological inputs. This suggests that each of the five features yields information related to the workload, and that a model including all these data would outperform the proposed classifier. Tests on larger samples should be performed before being able to draw conclusion on that specific point.

Also, a deeper (subject by subject) analysis of the results should be carried out, in order to check whether some combinations of specific physiological features are better for some categories of subjects (labile versus stabile for example). Since affects varied differently between subjects, this subject by subject analysis could also bring to light some relationships between physiological and psychological data, beyond workload.

ACKNOWLEDGMENT

This work was supported by Eurocopter.

The authors gratefully acknowledge C. Goulon and M. Huet for their help in building the graphical dynamic scene as well as C. Valot for fruitful discussions. They also acknowledge the students who have collaborated on the project and all the subjects who took part in the experiment.

REFERENCES

[1] R. Nikolova, V. Padev, and M. Vukov, "Functional determination of the operator state in the interaction of humans with automated systems," RTO/NATO, St Joseph Ottawa/Hull, Tech. Rep. RTO-MP-088, Oct 2003.

[2] A. Norcio and J. Stanley, "Adaptive human-computer interfaces: a literature survey and perspective," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, pp. 399–408, Apr 1989.

[3] P. A. Hancock and R. Parasuraman, "Human factors and safety in the design of intelligent vehicle-highway systems (IVHS)," *Journal of Safety Research*, vol. 23, pp. 181–198, 1992.

[4] W. Liao, W. Zhang, Z. Zhu, Q. Ji, and W. D. Gray, "Toward a decision-theoretic framework for affect recognition and user assistance," *International Journal of Human-Computer Studies*, vol. 64, no. 9, pp. 847 – 873, 2006.

[5] F. Di Nocera, M. Camilli, and M. Terenzi, "Using the distribution of eye fixations to assess pilots' mental workload," in *Human Factors and Ergonomics Society Annual Meeting October*, vol. 50, 2006, pp. 63–65.

[6] Y. Zhang, Y. Owechko, and J. Zhang, "Learning-based driver workload estimation," in *Computational Intelligence in Automotive Applications*, ser. Studies in Computational Intelligence, D. Prokhorov, Ed. Springer Berlin / Heidelberg, 2008, vol. 132, pp. 1–24.

[7] F. Tango and M. Botta, "Evaluation of distraction in a Driver-Vehicle-Environment framework: An application of different Data-Mining techniques," in *Advances in Data Mining. Applications and Theoretical Aspects*, P. Perner, Ed., vol. 5633. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 176–190.

[8] J. T. Cacioppo and L. G. Tassinary, "Inferring psychological significance from physiological signals." *American Psychologist*, vol. 45, pp. 16–28, 1990.

[9] G. B. Wallin and J. Fagius, "The sympathetic nervous system in man aspects derived from microelectrode recordings," *Trends in Neurosciences*, vol. 9, pp. 63–67, Jan 1986.

[10] J. A. Healey and R. W. Picard, "Detecting stress during Real-World driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, pp. 156–166, Jun 2005.

[11] G. F. Wilson and C. A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human Factors*, vol. 45, no. 4, pp. 635–643, 2003.

[12] J. A. Cannon, P. A. Krokhmal, R. V. Lenth, and R. Murphey, "An algorithm for online detection of temporal changes in operator cognitive state using real-time psychophysiological data," *Biomedical Signal Processing and Control*, vol. 5, pp. 229–236, Jul 2010.

[13] F. Putze, J. Jarvis, and T. Schultz, "Multimodal recognition of cognitive workload for multitasking in the car," in *20th International Conference on Pattern Recognition (ICPR)*. IEEE, Aug 2010, pp. 3748–3751.

[14] A. F. Sanders, "Towards a model of stress and human performance," *Acta Psychologica*, vol. 53, no. 1, pp. 61–97, 1983.

[15] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, pp. 18–37, Jan 2010.

[16] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in humanrobot interaction," *Pattern Analysis and Applications*, vol. 9, pp. 58–69, Apr 2006.

[17] J. D. Mayer, P. Salovey, and D. R. Caruso, "Emotional intelligence: Theory, findings, and implications," *Psychological Inquiry*, vol. 15, pp. 197–215, Jul 2004.

[18] A. M. Isen, "An influence of positive affect on decision making in complex situations: Theoretical issues with practical implications," *Journal of Consumer Psychology*, vol. 11, pp. 75–85, Jan 2001.

[19] T. F. O'Donnel,Robert, Eggemeier, "Workload assesment methodology," in *Handbook of perception and human performance*. New York NY: Wiley, 1986, vol. II, pp. 42.1–42.49.

[20] D. Watson, L. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The panas scales," *Journal of Personality and Social Psychology*, vol. 54, p. 10631070, 1988.

[21] P. Gaudreau, X. Sanchez, and J.-P. Blondin, "Positive and negative affective states in a performance-related setting. testing the factorial structure of the panas across two samples of french-canadian participants." *European Journal of Psychological Assessment*, vol. 22, pp. 240–249, 2006.

[22] Ice software. ISM, CNRS & Aix Marseille Université, Marseille, France. [Online]. Available: http://www.realite-virtuelle.univmed.fr/fr/presentation-crvm/plateforme-realite-virtuelle-crvm/systeme-informatique-crvm

[23] Captiv software. TEA. France. [Online]. Available: http://www.teaergo.com/index.php?lang=fr

[24] Flexcomp infinity hardware manual. Thought Technology Ltd. Montreal, Canada. [Online]. Available: http://www.thoughttechnology.com

[25] S. G. Hart and L. E. Staveland, "NASA task load index (TLX)," Human Performance Research Group NASA Ames Research Center, Moffett Field, California, Computerized Version v1.0, v. 1.0.

[26] P. Besson, V. Popovici, J. Vesin, J. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 63–73, January 2008, doi: 10.1109/TMM.2007.911302.

[27] J. L. Andreassi, *Human Behavior and Physiological Response*, 4th ed. Lawrence Erlbaum Associates, 2000.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, D. L. Schilling, Ed. John Wiley & Sons, 1991.

[29] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," PhD Thesis, University of California, Berkeley, USA, 2002.

[30] B. Cain, "A review of the mental workload literature," Defence Research and Development Canada Toronto Human System Integration Section, Toronto, Ontario,Canada, Technical Report RTO-TR-HFM-121-Part-II, Jul. 2007.

[31] T. F. Eggemeier and R. D. O'Donnel, "Workload assesment methodology," in *Handbook of perception and human performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. New York NY: Wiley, 1986, vol. II, pp. 42.1–42.49.

[32] P. Ekman, *Basic emotions*. Sussex, UK: John Wiley \& Sons, Ltd, 1999, ch. 3, pp. 45–58.