Bayesian Network classifiers inferring workload from physiological features: compared performance

P. Besson, E. Dousset, C. Bourdin, L. Bringoux, Tanguy Marqueste, D. R. Mestre, J. L. Vercher

Bayesian Network classifiers inferring workload from physiological features: compared performance

P. Besson, E. Dousset, C. Bourdin, L. Bringoux,
T. Marqueste, D. R. Mestre, J. L. Vercher

***Abstract*— This paper presents an approach based on Bayesian Networks to estimate the workload of operators. The models take as inputs the entropy of different number of physiological features, as well as a cognitive feature (reaction time to a secondary task). They output the workload variation of subjects involved in successive tasks demanding different levels of cognitive resources. The performances of the classifiers are 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

The ability to manage cognitive workload (denoted simply by *workload* from now on) during multitask activity is crucial for operators involved in driving complex engine such as car or aircraft. Intelligent systems can assist the operator in such situations, but for this assistance to be really efficient, it should be adapted to the current operator's workload. Thus, task demand should be decreased in case of overload, whereas more functions should be delegated to the operator in case of low workload [1], [2]. For example, in the context of car driving, where low workload is likely to result in a lack of vigilance, Advanced Driving Assistance Systems should be able to deactivate some functions (such as lateral and longitudinal control, speed regulation, etc.) to force the driver to focus on the task. Being able to characterize the operators workload is therefore the prerequisite to adaptive intelligent systems.

Computational models have been proposed to infer some cognitive states, such as workload or distraction, from task performance analyses or sensorimotor features (gaze, head movements, etc.) [3], [4], [5]. 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 drive to changes in physiological data [6], specifically (but not exclusively) when they are under the control of the autonomic nervous system. The latter is responsible for maintaining the body's homeostasis, noticeably through the orthosympathetic branch which mobilizes energy resources in response to the changing demands of the external and internal milieu [7]. Thus, electrocardiogram (ECG), electromyogram (EMG), skin conductance (SC), and respiration were used in [8] to infer the stress level by drivers using linear discriminant analysis. In [9], 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 [10] to derive an information-theoretic indicator of cognitive state. Support Vector Machine and ANN were applied to workload estimation in [11], using EEG, SC, respiration, and hear rate (HR) data. Notice that these models rely on physiological measurements to infer cognitive states but not necessary workload specifically (for example, stress, inferred in [8], should not be confused with workload thought it partly results from overload [12]).

In this work, our objective was to develop a task independent model, able to infer workload from objective measurements. We wanted to use non-intrusive and minimally invasive sensors, therefore, we did not measure EEG (incompatible with helmets wore by helicopter or fighter pilots for example) but restricted the measurements to EMG, ECG, SC and respiration. The reaction time (RT) to a secondary task was also used as a cognitive measure of workload [13], [1].

We propose Bayesian Network (BN) models that take as inputs the entropy of these physiological measurements and output the change of the subject's workload while they are involved in task demanding different levels of cognitive resources. Entropy of ECG signal has been shown to be a good indicator of distraction [14] but, to the best of our knowledge, it has never been applied to other physiological signals in a computational model of workload.

Different BN structures, built from expert knowledge, are tested, using in turn several combinations of physiological 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 change* between successive tasks. Indeed, in the context of defining adaptive intelligent systems, an erroneous prediction of the workload level that results in a false prediction in term of workload change (i.e. in a predicted variation opposite to the reality) should be

This work was supported by Eurocopter

All authors are with Institute of Movement Sciences, UMR CNRS 7287 & Aix-Marseille Université, Marseille, France patricia.besson@univ-amu.fr

absolutely avoided. It might drive the system to undertake actions opposite to those required by the operator's state, and have dramatic consequences.

The experimental protocol used to collect representative data and an analysis of these data are presented in sec. II. Sec. III describes the proposed models, whose 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, have participated to the experiment.

B. Material

The subjects sit in the dark, using a non-force feedback joystick and facing a standard 24" monitor, where the graphical dynamic flying scenes generated by the home-grown ICE software [15] were displayed. An experimenter's computer was used to acquire all the data synchronously, using the Captiv Software [16]. 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 [17]. The subjects bore stereo headphones, so that they could hear the 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

The subjects were asked to pilot a flying aircraft and to do their best to follow a trajectory made of 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 three first sessions (labeled *D1A0*, *D2A0* and *D3A0*), the subjects were presented with three different trajectories of increasing difficulty (D1, D2, and D3). The trajectories 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 two last 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 also required from the subjects. 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 (false and good detection rates; reaction times (RT)) tasks was recorded. The physiological variables comprised the following measurements:

- Hear Rate (HR) estimated from the ECG by the Captiv software, using the 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;
- Skin Conductance (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 evaluated their own workload during the performed task, using the NASA Task Load Index (TLX) scale [18]. The NASA TLX asks the subjects to rate their perceived workload on six different subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration). At the end of the experiment, these six components are matched two by two and the

[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.

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 TLX score per session.

E. Analysis of the experimental data

Prior to build a model that will take the collected data as input, it has to be checked that these data are representative of the problem at hand. That is, we have to ensure that the workload has been effectively manipulated using our experimental paradigm, so that variations observed in the physiological signals can effectively correspond to variations of workload.

The impact of the primary task difficulty on the 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 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 on all sessions are statistically different ($p < 0.001$) but for the sessions of the same difficulty levels (*D1A0* and *D1A1*, *D3A0* and *D3A1*). The difficulty also impacts the true positive (TP) detection rate of the secondary task (F(4,36)=8.84, $p < 0.01$), though the statistical difference only holds for D3 in the Student Neuwman-Keuls post-hoc analysis. Finally, the reaction times associated to these TP detections also differ significantly between the different difficulty levels (F(4,24)=14.084; p=0.0000) (but not between each session's trials (p=0.03)).

This statistical analysis establishes firstly that subjects behave with a coherent resource allocation strategy inside a given session. Secondly, as it is well-known that secondary task competes for the limited brain resources with the primary task (see e.g. [13], [1]), it indicates that the cognitive resources allocated by each subject on the primary task have increased at the expense of the secondary task. Nevertheless, these results do not insure that the overall level of involved resources has been increased: subjects might have been only partially committed in the task and have simply changed the resource allocation strategy as the primary task difficulty increased. Investigation of the subjective data (TLX scores) can give us some clues about this point. Generally speaking, the TLX scores increase with the session difficulty. An ANOVA on the TLX scores shows a significant difference between sessions (F(4,26)=12.284, p=0.000).

Table I summarizes the correlation values found for each subject between the TLX scores and the RT, per session. In most cases, the correlation is greater than 0.5. which indicates that the subjective evaluation of workload is consistent with the objective measure (RT) and that subjects committed in the task. Values below 0.5 are not necessarily due to a mismatch between subjective and objective metrics: some subjects experiencing high workload totally gave up on the secondary task, so that RT were not available[2] (thus, the relation with TLX scores is not linear anymore).

TABLE I
CORRELATION BETWEEN THE TLX SCORES AND THE RT OVER THE 5 SESSIONS, FOR EACH SUBJECT

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
r	0.67	0.79	0.83	0.80	0.68	0.48	0.32	0.75	0.90	0.45

Finally, this 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.

III. MODEL

A. Selection of output and input features

A data driven approach to modeling problem requires to prealably 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 [19].

We have decided to use the Shannon's entropies of the physiological data as inputs for our model. The entropy of a random variable (rv) X is a measure of the average uncertainty in X [20]. Stated in a different and simpler way, it is a measure of disorder. As such, it is likely to capture differences in physiological data related to workload variations.

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 three first 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 three first 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 TLX scores (see Fig. 2 showing the subject 4's features as an illustrative example). However, we observed the variation of the physiological data

[2]In that case, an arbitrary mean RT value of 1.5sec has been used in the correlation computation.

to be idiosyncratic: for some subjects, the mean entropy values of some physiological data might decrease with difficulty levels, whereas they decrease for other subjects. As a results, the models will be individual (trained and tested on each subject separately).

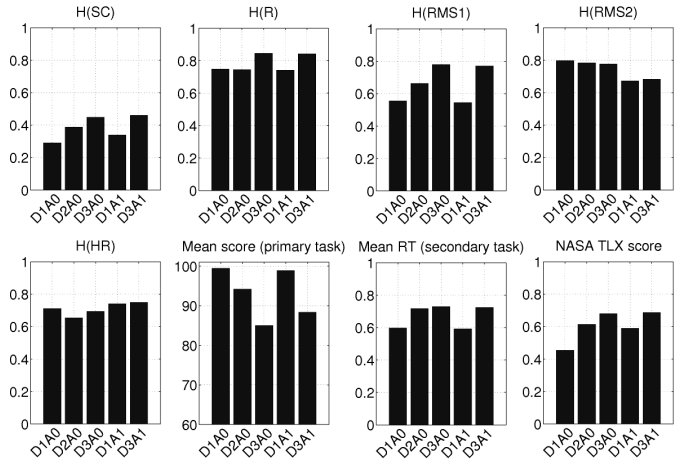

Fig. 2. Mean physiological feature values (entropy values of the physiological signals, in bit), performance on the primary task (in %), reaction times (RT, in sec) and NASA Task Load Index (TLX) scores for each session performed by subject 4.

B. Model definition

The model aims at inferring the subject's TLX score on each session, from the physiological features SC, R, HR, RMS1 and RMS2[3]. To this end, different BN classifiers are tested, each taking one, two or three of the possible physiological features as inputs. As a result, 25 classifiers with different physiological nodes are trained and tested. Moreover, three BN structures are tested. Structure 1 is a naive BN were the TLX is a direct child of the physiological nodes. Structure 2 is also a naive BN but TLX is now a child of the RT, which is itself a child of the physiological nodes. Structure 3 has a more complex structure, where TLX is a direct child of both the physiological nodes and of RT. The different structures are presented on Fig. 3 for a 2-node classifier made of physiological features Φ_1 and Φ_2.

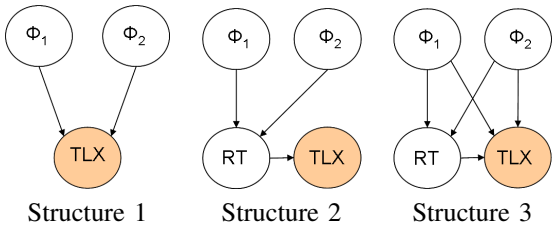

Fig. 3. BN models inferring the TLX scores from physiological features Φ_1 and Φ_2 either directly (Structure 1), via RT (Structure 2), or from both RT and the physiological inputs (Structure 3). There can also be 1 or 3 physiological nodes.

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

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 TLX, 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 three first 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 [21]. 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).

C. Assessing the model performance

The performance of the models is assessed by looking at the differences in the TLX scores between the *D1A1* and the *D3A1* sessions. The model output will be deemed as correct if the observed and inferred TLX scores 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' TLX scores on sessions *D3A1* and *D3A1*, and Δ^* the difference between the predicted TLX scores 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 false model's 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 performance, 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 the Fig. 4, it can be observed that most of the models are performant in either one of the two *diversity* (S) or

TABLE II

BEST PERFORMANCE IN TERMS OF ACCURACY AND DIVERSITY INDEPENDENTLY

Performance criterion	Structure	Physiological nodes	S	θ
Maximal Accuracy	1	SC	20	0.89
Maximal Diversity	2	HR and SC	80	0.47

accuracy (θ) criteria. Table II presents the best classifiers in term of a single performance criterion solely. However, we are interested in classifiers performant in both dimensions simultaneously. The model that gives the best performance in term of both accuracy and diversity, namely, $S = 60\%$ and $\theta = 0.66$, is the three physiological node classifier $HR; RMS_2; SC$ with the structure 2. Its performance is plotted as a function of its diversity in Fig. 5.

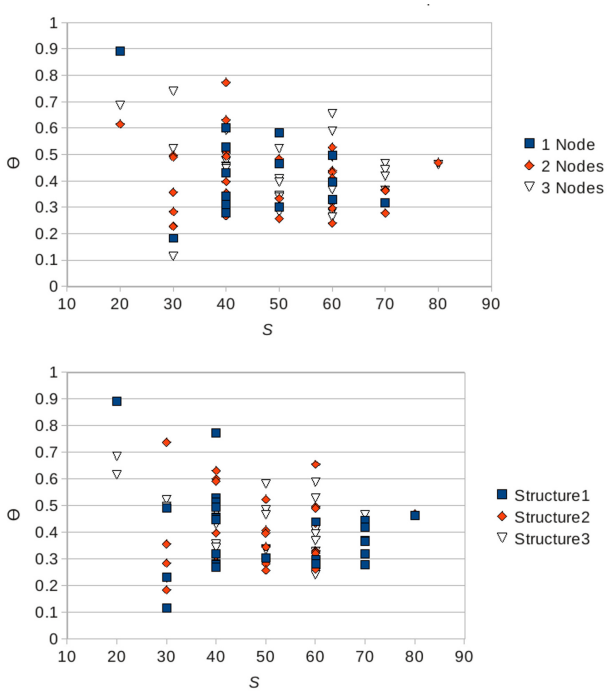

Fig. 4. Performance of the models in terms of diversity (S) and accuracy (θ), depending on the number of nodes (top) or on the structure (bottom) of the models. The best models lie in the upper right-hand side quarter of the graphes.

To analyse the impact of the number of nodes and of the structure on models' performance, let us look at the best models, i.e., the models with a performance greater than 50% over the sets for both diversity and accuracy criteria (these are the classifiers lying in the upper right hand side quarter of Fig. 4). Indeed, the mean performance over the model sets do not tell us whether good models for one criterion are also good models for the other. The results are presented on Fig. 6. Generally speaking, the percentage of models fulfilling the required "best performance" criterion is

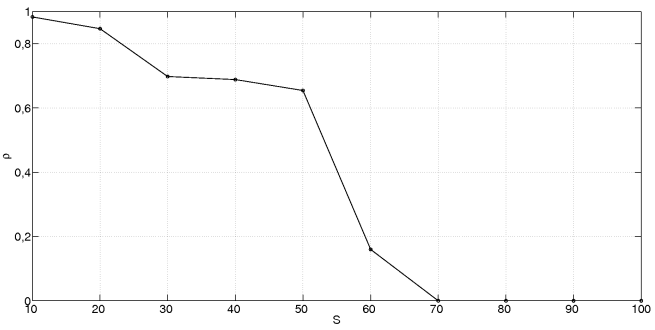

Fig. 5. Performance of the best classifier ($HR; RMS_2; SC$ with the structure 2).

not high (less than 15%). It is null for the models with the simplest structure (structure 1), but it increases as soon as RT is added to the model (structures 2 and 3), with the best performance being obtained with structure 3 (12%). When comparing the models on the basis of their node number (whatever the structure), the use of two nodes leads to the smallest percentage of good classifiers (thought the best results are obtained with a 2-node classifier), whereas the largest number of good classifiers are obtained when using three nodes.

Notice that each of the five physiological features appears in one of these good classifiers. This certainly indicates that none of these features is specific enough of the workload change. We can do the hypothesis that a classifier with five physiological nodes would outperform the performance of the classifier proposed in this work. However, since we are training subject-dependent models, our sample sizes were too small to deal with a classifier with 5 physiological nodes, so that we were not able to check this hypothesis in this study.

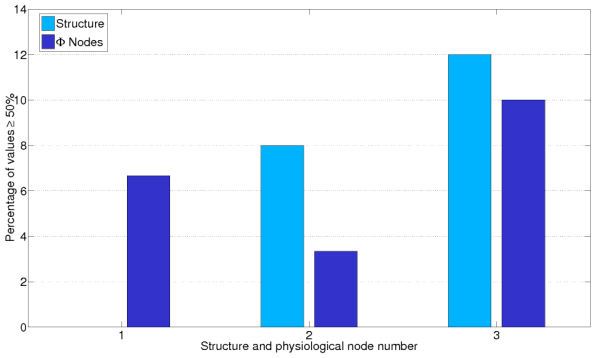

Fig. 6. Percentage of good models (with performance greater than 50% for both accuracy and diversity criteria) for the different structures and the different physiological node numbers.

V. CONCLUSIONS AND FUTURE WORK

In this paper, Bayesian networks are proposed to infer the variation of the workload for operators involved in multitask activities, from physiological and cognitive (RT) measurements. The advantage of physiological measurements is that

their interpretation if more independent from a specific operator task than sensorimotor features for example. Entropy based features are firstly derived from the raw measurements so that the models receive inputs more specific from the studied phenomenon. Different structures and number of inputs are tested and compared in term of two criteria to be jointly optimized: the accuracy and the diversity of the classifier.

The best model is the three physiological node classifier $HR; RMS_2; SC$ with the structure 2 (workload inferred from the physiological nodes, via the RT). However, a finer analysis of the model performances points out that each of the five proposed physiological features might appear in classifiers with good performance. Also, the performances increase with the number of physiological inputs in the model. This suggests that each of the five features yields information related to the workload, and that a model including all these information would outperform the proposed classifier. Tests on larger samples should be performed for being able to draw conclusion on that specific point.

Though the structure of the best model is the structure 2, the relative number of good classifiers is more important with the third tested structure (where the workload rv is a child of both the physiological nodes and of the RT). This indicates that physiological and cognitive features carry complementary information about the subject's cognitive state, extracted and used by the models. Including the RT in the model yields better workload prediction, at the expense of a slightly more task-dependent method, since it requires a secondary task to be performed. However, there are a lot of situations where routine tasks can be used to infer RT values.

These are only preliminary results and refinements should be brought to the models, as well as tests on larger sample sizes (which should result in improved models' performance). It should be checked whether the best classifier $HR; RMS_2; SC$, which shows a good diversity performance, remains performant when tested on new subjects. It is also possible that some combinations of specific physiological features are better for some categories of subjects (labile versus stabile for example). This could be checked by a deeper (subject by subject) analysis of the models' performance.

VI. ACKNOWLEDGMENTS

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 that took part in the experiment.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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, July 2010.

[11] 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.

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

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

[14] L. Yu, X. Sun, and K. Zhang, "Driving distraction analysis by ECG signals: An entropy analysis," in *Internationalization, Design and Global Development*, P. L. P. Rau, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6775, pp. 258–264.

[15] 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

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

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

[18] 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.

[19] 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.

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

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