Multimodal estimation of a driver’s spontaneous irritation

Lucas Malta, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda

***Abstract*— In this paper we present our latest achievements in the continuous estimation of a driver’s spontaneous irritation. Experiments are conducted with data from 20 drivers, recorded under real driving conditions. While driving, participants also interact with a speech dialogue system to retrieve and play music. A fusion method is proposed to integrate information on the driving environment, driver behavior, driver’s physiological state, and speech recognition results. Overall, we are able to correctly detect 80% (true positive rate) of the irritation, and, when drivers are not irritated, we only make mistakes 9% of the time (false positive rate). Results also support the relevance of gas- and brake-pedal operation as well as speech recognition results in irritation estimation.**

I. Introduction

In vehicle traffic, where most participants remain anonymous, interactions are short, and communication is restricted, misinterpretations and misunderstandings are commonplace actions, likely to result in irritation and aggressive behaviors with enormous impact on society [1], [2]. Anger while driving may interfere with perception, information processing, and motor performance, increasing therefore, the likelihood of an accident [2]. Given the fact that we are spending more and more time in the automobile, in this research we chose to explore drivers’ irritation recognition under real-world conditions—an open and highly relevant question that needs to be addressed without delay. The interpretation of a driver’s current affective state is a key point for the development of intelligent in-vehicle interfaces, which help enhancing the driving experience, without contributing to performance degradation.

Very few attempts have been made to recognize affect displays in an in-car environment. All of them still suffer from one or more limitations of most emotion recognition methods, such as context-insensitivity, use of acted or carefully elicited affective expressions, and pre-segmentation of emotion sequences [3]. The fact that experiments in virtually all existing approaches are performed under “safe” laboratory conditions (e.g. in driving simulators) further restricts the applicability of previous results in real world, since studies have shown that drivers react differently in the constrained lab environment [4].

Single and multimodal approaches using different types of data have been proposed. In [5], authors recognized four emotional states based on physiological signals: euphoria, disappointment, high, and low stress. Data from ten drivers were collected under laboratory conditions. The overall classification rate was 79.3% using Support Vector Machines (SVM). A speech-based approach was used in [6] for the classification of four emotional states: anger, confusion, joy, and neutrality. Data from ten subjects were collected while they drove a driving simulator. Overall accuracy of 77.8% was achieved also using SVM. A bimodal fusion of acoustic and visual information for emotion recognition in an automotive environment was proposed in [7]. Deliberate affective behavior from seven non-professional actors was recorded in a vehicle standing idle. An average recognition rate of 90.7% for the fusion approach was achieved with SVM. An excellent survey of state of the art affect recognition methods can be found in [3].

Questions addressed by previous studies in this field were not particularly challenging, and many limitations still need to be overcome. In this research we propose a context-dependent multimodal data fusion technique for the continuous recognition of spontaneous irritation. Data were collected under real-world conditions, while drivers interacted with an automatic speech recognition system, so that not only the traffic but also the man-machine interaction could be regarded as sources of irritation. The approach presented here is an extension of a pilot study described in [8], where the irritation estimation framework was outlined and results for three different drivers were presented. Here, we extend the previous work by introducing a new feature (speech recognition results), by investigating the effectiveness of driving behavior and speech recognition in irritation estimation, and by evaluating the system using data from 20 drivers. Data collection is described in section II and context representation in section III. We then offer a description of the fusion technique in section IV, and, finally, results and conclusions in section VI.

II. Data Collection

A data collection vehicle was designed for synchronously recording audio with other multimedia data. Various sensors were mounted on a Toyota Hybrid Estima with 2,360 cc displacement and automatic transmission. Video footage, driving behavior, and physiological signals were recorded synchronously with audio under both driving and idling conditions. Signals were further down-sampled to 10Hz.

30 participants (20 male, 10 female) took part in the experiment. They were, on average, 31 years old (range 20-58 years) and had held a driver’s license for a mean period of 11.4 years (range 1.4-39 years). Participants drove on city streets in the city of Nagoya, Japan. During the experiment, they were often forced to avoid parked vehicles, bicycles, and pedestrians. They also encountered rapid traffic-density changes and long stops at red-light signals. The course participants followed was pre-selected so that spontaneous

The authors are with the Grad. School of Information Science, Nagoya University, JAPAN. {malta, miyajima, kitaoka, takeda}@sp.m.is.nagoya-u.ac.jp. This research is supported by Strategic Information and Communications R&D Promotion Program of MIC Japan.

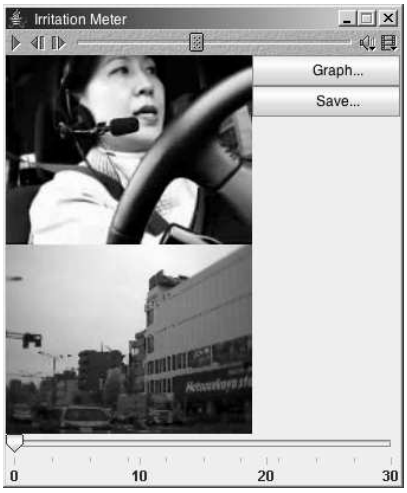

Fig. 1. Interface designed for irritation assessment.

irritation had a greater chance to be elicited. Experiments lasted from 6 to 12 minutes, depending on the traffic conditions. In addition to that, using an automatic speech recognition (ASR) system, drivers retrieved and played songs while driving from a list of 635 titles from 248 artists. Music could be retrieved by artist name or song title, e.g., "Beatles" or "Yesterday."

After each experiment, the participant was also asked to assess his/her subjective level of irritation by referring to the front-view and facial videos as well as the corresponding audio. A user interface for such assessment, shown in Fig. 1, was designed so that drivers could slide a bar from 0 to 30, i.e., from normal condition to highest irritation. The level of irritation was stored every 0.1 seconds. Irritation was further quantized into two levels: irritated and not irritated. The optimal threshold for quantizing irritation was defined experimentally, as described in section V.

III. Data Transcription

One of the major challenges for the emotion recognition field is the development of context-sensitive methods that also take into account the situation in which the affective behavior was elicited, rather then relying solely on responses (e.g. speech, facial expressions, physiological changes). While driving, our actions are, most of the time, carefully planned after a complex cognitive decision-making process based on different variables such as weather condition, road design, and the presence of pedestrians. Therefore, an effective labeling of multimodal information is critical for providing a more meaningful description of situations drivers experienced.

In this research, we propose a data transcription protocol that covers most of the factors that might affect drivers and the drivers' responses. Labels in the transcription were selected in cooperation with the University of Texas at Dallas, so algorithms can be validated in data recorded in different locations from drivers with different profiles. The transcription is comprised of six major groups: driver's affective state (level of irritation), driver actions (e.g. facial expression), driver's secondary task, driving environment (e.g. type of road, traffic density), vehicle status (e.g. turning, stopped), and speech / background noise.

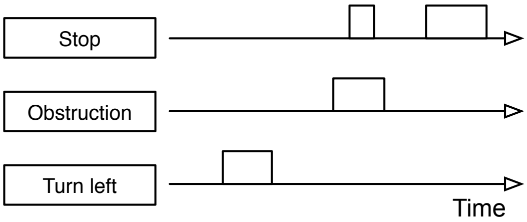

Fig. 2. Example of transcription labels.

Data transcription from all 30 subjects was manually carried out by seven annotators who were allowed to utilize only frontal video. No audio was provided to avoid bias when labeling overall face data from the video. Annotators were graduate students who had volunteered for the experiment. Not all of the proposed labels were utilized: Those used were pre-selected for the specific task of estimating irritation. Based on scenes of high levels of irritation, the selected labels were:

1) Overall face (positive and negative facial expressions were grouped together, so that this binary feature indicates deviations from the neutral face);
2) Turn (including left/right turns and turning with interruptions);
3) Curve (including left/right curves);
4) Obstructions caused by pedestrians, bicycles, and parked vehicles;
5) Traffic density;
6) Stops at red-light signal.

The process of transcription is akin to annotating the time span of labels, so transcription results can be seen as multiple streams of binary information, as shown in Fig. 2.

IV. Data Fusion

A multimodal analysis of spontaneous affective behaviors is crucial for improving recognition performance. The use of complementary information from various channels has proved to be superior to its single-modal counterpart, since the uncertainty due to one channel can be decreased by adding new information. In this study, together with the information obtained from transcription labels, the following data are combined:

- *Electodermal activity (EDA).* Electrodermal activity is one of the most widely used response systems in the literature. It is linked with psychological concepts of emotion, arousal, and attention [9]. We have been collecting EDA through a skin potential sensor, which measures the potential difference between two points on the skin.
- *Driving behavior.* The investigation on the effects of different emotional states on the way we drive is an open and very interesting question. This study is the first attempt to tackle this problem from an engineering

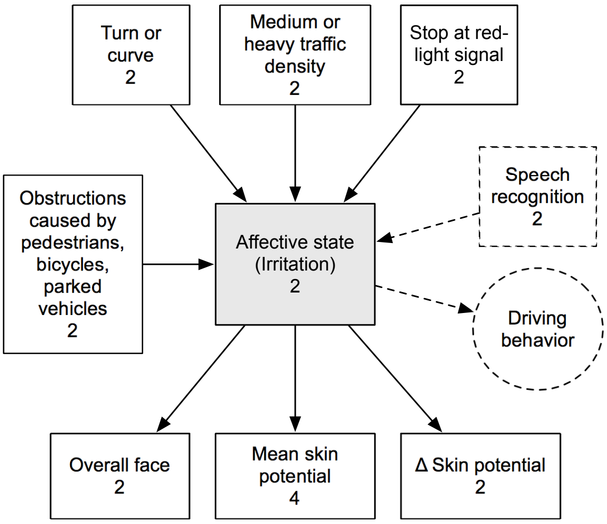
Turn or curve 2
Medium or heavy traffic density 2
Stop at red-light signal 2
Obstructions caused by pedestrians, bicycles, parked vehicles 2
Speech recognition 2
Affective state (Irritation) 2
Driving behavior
Overall face 2
Mean skin potential 4
Δ Skin potential 2

Fig. 3. Proposed Bayesian network structure. Squares represent discrete (tabular) nodes and circle represents a continuous (Gaussian) node. Numbers represent the number of mutually exclusive states each node can assume. Experiments with and without speech recognition and driving behavior nodes were performed.

point of view by trying to show that pedal operation is also affected by different affective states. The force signals from gas and brake pedals were used as driving behavior.

- *Speech recognition results.* Another important factor we understood to be critical for spontaneous irritation recognition is the interaction with automatic speech recognition (ASR) systems. Most ASR systems are sensitive to noise, which, inside the vehicle, comes from different sources, such as wind, engine, and traffic. Although ASR systems have improved a lot, we still do not have an error-free device, and frequent errors are a common cause of discomfort to drivers.

The process that causes irritation is complex. Several uncertainties might be present in this process, and while driving, irritation can be regarded as the result of a wide range of contextual variables. To effectively estimate irritation, a system that integrates evidence from multiple sources in an efficient language is needed, and a Bayesian network (BN) is the natural choice to deal with such task. A BN is a state-of-art knowledge representation that creates a very efficient language for building models of domains with inherent uncertainty. Joint probabilities of a set of continuous or discrete random variables (nodes) are represented in a BN, which also explicitly encodes conditional independence assumption in its structure [10].

One of the main goals of a BN is to infer the state of a given unobserved variable, given the state of observed ones. In our case, we want to infer irritation given the context, speech recognition results, and driver's responses. During the learning stage, all nodes were filled with information, while during the test stage, the irritation node was empty and its state was inferred given the state of all other nodes, which were observed. Detailed information on network parametrization can be found in [11]. In experiments, the Bayes Net Toolbox for Matlab, freely available at [12] was used. Figure 3 shows our BN designed to integrate transcription labels, driving behavior, physiological state, and speech recognition results. In order to verify the effectiveness of driving behavior and speech recognition results as features, experiments were performed with four different configurations of this network: (1) Without neither driving behavior nor speech recognition nodes (baseline network); (2) The baseline plus driving behavior node (driving behavior network); (3) The baseline plus speech recognition node (speech network); and (4) With all nodes (full network).

For representing the physiological state of drivers, we used the mean of normalized signal (mean skin potential) and the absolute value of the first-order difference of the normalized signal (Δ skin potential) [8]. For the network's driving behavior node, features were extracted through spectral analysis of the gas and brake pedal signals by using a special feature called "cepstrum" (cepstral coefficients): a widely used spectral feature for speech and speaker recognition, and, more recently, it proved to be effective in driver modeling [13]. Therefore, we decided to verify the effectiveness of cepstrum in affect recognition. Ceptrum is defined as the inverse Fourier transform of the short-term log-power spectrum and is obtained as follows:

$$c(m) = \frac{1}{M}\sum_{k=0}^{M-1} \log|X(k)|e^{2\pi kmj/M}, \qquad m = 0, 1, ..., M-1. \tag{1}$$

where $X(k)$ denotes the M-point discrete Fourier transform of the windowed signal $x(n)$. Before calculating the Cepstrum, gas and brake were combined together by setting the pedal operation signal to $gas - brake$. This is important to avoid calculating cepstral coefficients of frames when the pedal was not being pressed. The time derivative of the cepstral coefficients was also utilized as feature, since the dynamics of pedal operation is also important.

As for the speech, the most natural feature is the presence of recognition errors. There is, however, a great drawback in using recognition errors, since a manual transcription of speech must be completed in advance, so one can compare recognition results with the actual speech. This is a very expensive and time-consuming process, which can not be performed in real-time. Therefore, we decided to adopt another strategy: after the speech recognizer mistakes the name of an artist or song, a very natural reaction from the driver is to say "No", so that he/she can repeat the desired input until the machine gets it right. The instant the ASR system recognized a driver utterance as "No" can be then used as a feature. Nevertheless, since this is a pinpoint label, an enlargement of its boundaries is necessary. After an analysis of irritation videos, we decided to add five seconds before and 15 seconds after each utterance recognized as "No", so that the speech recognition results feature is represented by a 20-second window.

V. Experiments and Evaluation

Experiments were performed with data from 13 male and 7 female drivers. They were, on average, 29 years old (range 20-46 years) and had held a driver's license for a mean period of 10.4 years (range 1.4-27 years). The irritation of ten of the original 30 subjects was insignificant, so their data were not used. Individual networks were trained using 50 to 70% of data from each participant.

When no speech recognition nor driving behavior was introduced in the network (baseline), we conducted experiments with different values of skin potential signal frame length (1.6 s, 3.2 s, 6.4 s, 12.8 s, and 16.0 s), threshold for quantizing the Δ skin potential into two steps (0.1, 0.2, 0.3, 0.4, 0.5), and threshold for quantizing the irritation signal into two steps (3, 4, 5, 6). After introducing driving behavior to the baseline, we did experiments with different values of driving behavior signal frame length (0.8 s, 1.6 s, 3.2 s, 6.4 s, 12.8 s). The number of cepstral coefficients was set to one ($c(0) + c(1)$). The frame shift was fixed at 0.5 s for all signals. Estimated values obtained from the unobserved node *Affective state* were used as the output of the system, i.e., the estimated irritation.

We evaluated the capacity of the proposed system to detect irritation. After calculating the estimation signal from each driver, it was filtered using a median filter of twelve seconds so that spikes and short gaps could be removed. In order to estimate the overall detection effectiveness, we added together true/false positives/negatives from all drivers, so that we could calculate overall true and false positive rates, represented by a single point in the receiver operating characteristic (ROC) space.

VI. Results and Conclusions

Overall results for all four network configurations are shown in Fig.4. In the ROC space, the point (0,1) represents the perfect estimation. The closer the result gets to this point, the better. Circles centered in (0,1) are also shown so that different results can be easily compared. The full network configuration, when both driving behavior and speech recognition results were used, achieved the best result: a true positive (TP) rate of 80% and a false positive (FP) rate of 9%, i.e., we were able to correctly detect 80% of the irritation, and, when drivers were not irritated, we only made mistakes 9% of the time. The baseline configuration achieved the worse result, with a TP rate of 74% and a FP rate of 12%. Both of the proposed features were effective in boosting the estimation. These are a very encouraging results, especially considering the challenging nature of the task. Moreover, irritation was recognized continuously, i.e., no pre-segmentation of data sequence was used. This is a rather complex problem, given the fact that boundaries are often very fuzzy.

Optimal estimation parameters were: skin potential signal frame length of 12.8 s, threshold for quantizing the Δ skin potential of 0.3, and threshold for quantizing the irritation signal of 5. The optimal driving behavior signal frame length was 1.6 s.

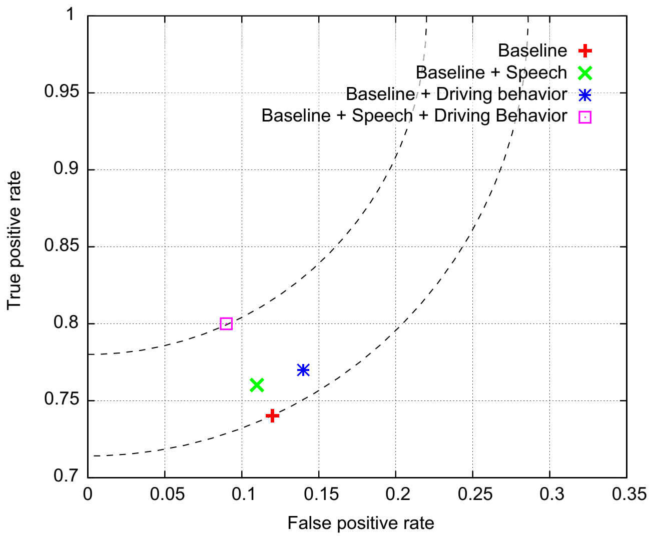

Fig. 4. Overall results achieve by the network in all four different configurations. Dashed lines are part of circles centered at (0,1).

Data annotation is still performed manually. If it could be done automatically in real-time, based on information from the infrastructure and vehicular sensors, our system would be able to detect irritation with a delay of around 18.8 s. This is the length of the largest data frame we need to process plus half the length of the median filter. The time required for calculating features and for the BN to receive a new evidence and perform inference is negligible. According to experts in the automotive industry, 30 s is the maximum acceptable delay, so we would still be inside this limit. Training an individual network takes, on average, 0.5 s on a Quad-Core AMD Opteron(tm) Processor 2356, 16GB memory.

The main contributions of this study are: (1) Evaluation of proposed method using data collected under real driving conditions; (2) A driving data transcription protocol, which can be used in a wide range of studies; (3) Successful introduction of speech recognition as a feature; and (4) Demonstration that driving behavior is affected by different affective states. As a future work, we intend to investigate the temporal relationship among variables, examine other speech features, and develop a driver-independent model, that is, a model trained with data from all drivers together.

References

[1] T. E. Galovski and E. B. Blanchard, "Road rage: a domain for psychological intervention?" *Aggression and Violent Behavior*, vol. 9, no. 2, pp. 105–127, 2004.

[2] G. M. Björklund, "Driver irritation and aggressive behaviour," *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 1069–1077, 2008.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39–58, 2009.

[4] L. Angell *et al.*, "Driver workload metrics project task 2, final report," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 810 635, 2006.

[5] C. Katsis *et al.*, "Toward emotion recognition in car-racing drivers: A biosignal processing approach," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 38, no. 3, pp. 502–512, 2008.

[6] B. Schuller, M. Lang, and G. Rigoll, "Recognition of spontaneous emotions by speech within automotive environment," *Jahrestagung für Akustik (DAGA)*, vol. 32, pp. 57–58, 2006.

[7] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 2, pp. 1085–1088, 2005.

[8] L. Malta, P. Angkititrakul, C. Miyajima, and K. Takeda, "Multi-modal real-world driving data collection, transcription, and integration using Bayesian network," in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 150–155.

[9] J. T. Cacioppo and L. G. Tassinary, *Principles of Psychophysiology: Physical, Social and Inferential Element*, J. T. Cacioppo and L. G. Tassinary, Eds. Cambridge University Press, 1990.

[10] F. V. Jensen, *Bayesian networks and decision graphs*, M. Jordan, Ed. Springer, 2001.

[11] K. P. Murphy, "Inference and learning in hybrid Bayesian networks," University of California, Tech. Rep. CSD-98-990, 1998. [Online]. Available: citeseer.ist.psu.edu/murphy98inference.html

[12] K. Murphy, "Bayes Net Toolbox for Matlab," 2007. [Online]. Available: http://bnt.sourceforge.net/

[13] C. Miyajima *et al.*, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 427–437, 2007.