2008－SLP－74 (42)
2008 / 12 / 10

Driver's irritation detection using speech recogntion results

Lucas Malta, Chiyomi Miyajima, Akira Ozaki, Norihide Kitaoka, and Kazuya Takeda
Graduate School of Information Science, Nagoya University, JAPAN

Abstract—**In this work we present our efforts towards the multi-modal estimation of a driver's affective state under naturalistic conditions. Multi-modal data from 18 subjects (2.3h) who interacted with an automatic speech recognition system while driving are recorded. A transcription protocol is designed to provide a meaningful description of the driving environment. A data fusion model based on Bayesian network is proposed and used for estimating a driver's level of irritation. Information on transcription labels, physiological signals, driving behavior and speech recognition are integrated. Preliminary results are very encouraging.**

I. Introduction

The assessment of a driver's affective state becomes increasingly important the more we understand the effect negative emotions have on the way drivers behave. Emotions such as anger often lead to openly aggressive actions, degrade information processing, and increase the likelihood of an accident [1]. 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.

There are only a few studies on driver's affective state estimation and due to complications and costs imposed by multi-modal driving data collection, the vast majority rely on controlled laboratory settings, e.g., driving simulators. Besides, given the richness of information available from the driver and the driving environment, the use of knowledge from a single modality can also be regarded as a drawback of current approaches. For example, in [6], authors recognized four emotional states—euphoria, disappointment, high, and low stress—based on physiological signals. Data from ten drivers were collected under laboratory conditions and emotional states annotated by coders. Overall classification rate was 79.3% using support vector machines (SVM). In [11], authors relied on speech 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. In [7], a facial expression-based emotion recognizer was proposed. Six different types of facial expressions were acted while subjects drove a real vehicle in a cloudy day. Recognition accuracies above 90% were reported, but experimental conditions were described poorly and no information whatsoever on the number and characteristic of subjects was provided. In [4], authors recognized independently four emotion primitives—valence (positive vs. negative), activation (calm vs. excited), and dominance (weak vs. strong)—from clean speech superimposed to vehicle noise. The clean signal was obtained from utterances recorded during a talk-show. The noise was recorded under various driving conditions. Error rates of 13%, 16%, and 15% were achieved for valence, activation, and domination, respectively.

Reasonable recognition rates have been achieved under controlled setups, but in order to ensure that results will be applicable in real-world, algorithms still must be validated under real driving conditions, which pose further difficulties on the emotion recognition task. Not only the severe noise coming from different sources, but also the dynamical driving environment that continuously affect the way drivers behave, may contribute to degradation of algorithms validated under laboratory or controlled naturalistic settings. Results based on acted or carefully elicited emotions, although useful to a certain extent, are also hardly repeatable under real-world conditions. Consequently, the estimation of a driver's affective state in naturalistic environments poses several challenges that still need to be met before significant improvements in driving comfort can be observed.

In this research we focused on developing a system for estimation of a driver's level of irritation. Our system's design overcomes three crucial obstacles that limit the use of current approaches to driver's state estimation: (1) the use of driving simulator data in experiments; (2) inaccurate annotation of driver's state; (3) the disregard for information on the driving context. Some of these limitations affect not only systems aimed at estimating a driver's affective state, but also systems focused on detecting conditions such as stress or high workload.

II. Materials and Methods

A. Data Collection and Transcription

To overcome the first limitation, we used in experiments multi-modal real-world driving data, recorded from 18 drivers (total of 2.3h). Video footage, driving behavior, and physiological signals were recorded synchronously with audio in a vehicle under both driving and idling conditions. Participants drove on city streets in the city of Nagoya, Japan. During the experiment, drivers interacted with a spoken dialogue system to retrieve and play songs from a list of 635 titles from 248 artists. Music can be retrieved by artist name or song title, e.g., "Beatles" or "Yesterday." All subjects are trained in the proper method for performing the retrieval prior to the start of the experiment. Experimental setup was designed so that not only the traffic, but also the human-machine interaction could be regarded as sources of irritation. Detailed information on the recording apparatus can be found in [10].

Given a real-world driving database, the problem of how to interpret efficiently multi-modal information naturally arises. While driving, our actions are, most of the time, carefully planned after a complex cognitive decision-making process that takes into account different variables such as weather conditions, road structure, and traffic density. An effective transcription of multi-modal driving data is essential for providing a meaningful description of traffic situations. Assessing a driver's state without fully understanding the environment might lead to wrong conclusions. For example, an assessment, based on physiological signals, of a driver's stress levels due to a secondary task is in essence inaccurate because the causes of stress cannot be fully explained.

In this work, we proposed a transcription protocol for multi-modal driving data that takes into account variables in 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. Data from all drivers were tagged by six coders using video footage. Transcription reliability was verified using Cohen's kappa κ [3].

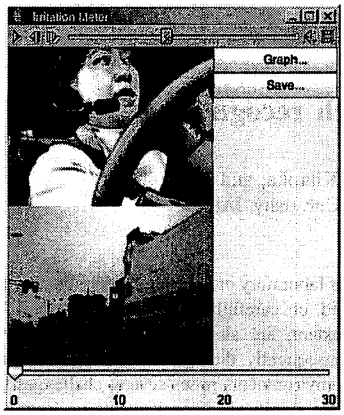

Fig. 1. Interface used for assessing level of irritation.

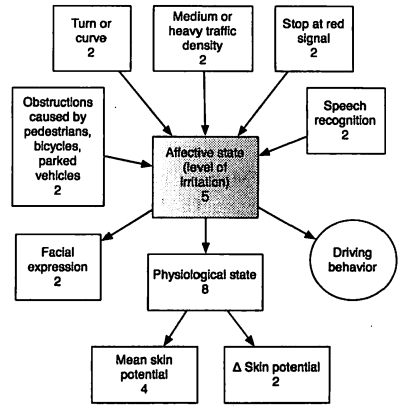

Fig. 2. BN structure used in experiments.

The second limitation of current approaches to affective state estimation has to do with the way emotions are labeled on a time series. Analysis of questionnaires subsequent to experiments and subjective labeling by trained coders have been widely used as labeling strategies. Although effective to a certain extent, these procedures may fail to either represent the actual emotional intensity or to precisely locate emotions on a timeline. In this research we proposed a new labeling strategy: after each experiment, the driver was asked to assess his/her 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 normal condition to highest irritation. Resulting signal is a continuous metric of driver's state and can be used to model dynamical variations in emotion.

B. Data Fusion

The third limitation of current approaches concerns the type of information frequently used in driver's state estimation. While driving, emotions can be regarded as the result of a wide range of contextual variables. For example, irritation can be related to factors such as high-density traffic, long waits at red stop signals, and frequent obstructions of the vehicle's path by pedestrians. Therefore, a system that takes into account not only drivers reactions, such as speech, facial, and physiological changes, but also the environment that may have caused these reactions, is highly desirable. In this research, we proposed a mathematical model that explicitly explains driver's status based on both his/her responses and also on the driving context. Apart from the level of irritation, assessed by drivers, other transcription labels, and driver's physiological state (skin potential), in this work we introduce two new features for affective state estimation: driving behavior (brake and gas pedals operation) and speech recognition. The last is a binary feature, which indicates utterances recognized by the machine as "No", which is a good indicator of communication problems. We refrained from using speech recognition accuracy because it requires manual speech transcription, which is costly and can not be done in real-time.

In order to effectively estimate irritation, a model 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 Bayesian Network (BN) is a knowledge representation that creates a very efficient language for building models of domains with inherent uncertainty. A BN consists of a set of variables and a set of directed edges between variables. Each variable has a finite set of mutually exclusive states. These variables, together with their directed edges, form a directed acyclic graph (DAG). To each variable A with parents $Pa(A_1), ..., Pa(A_n)$, there is attached the potential table $Pr(A|Pa(A_1), ..., Pa(A_n))$ [5]. Figure 2 shows our BN designed to integrate transcription labels, driving behavior, physiological signals, and speech recognition. Squares represent discrete (tabular) nodes and circle represents a continuous (Gaussian) node. Numbers represent the number of mutually exclusive states each node can assume.

Conditional probabilities in the network were learned from data. Individual networks were trained using 50 to 70% of data from each driver. The remaining portion was used for testing. Expectation Maximization (EM) algorithm was used to calculated Maximum Likelihood Estimates (MLEs) of the parameters. If A_i is a discrete node, the parameter vector is $\theta_{ijk} = Pr(A_i = k|Pa(A_i) = j)$, which is just a table of numbers (conditional probability table). In the discrete case, the sufficient statistics are N_{ijk}, the number of times the event $(A_i = k, Pa(A_i) = j)$ occurs in the training set. Since

$$\theta_{ijk} = \frac{Pr(A_i = k, Pa(A_i) = j)}{Pr(Pa(A_i) = j)} \approx \frac{\frac{1}{N}N_{ijk}}{\frac{1}{N}N_{ij}}, \qquad (1)$$

where $N_{ij} = \sum_k N_{ijk}$, the MLE is $\widehat{\theta_{ijk}} = N_{ijk}/N_{ij}$. For continuous nodes, on the other hand, means and covariance matrices are calculated from training data. ML approach can give severely over-fitted results for small data-sets, so as a future work, we plan to investigate the introduction of a prior distribution over the parameters. Detailed learning process can be found in [9].

C. Feature Extraction

For representing the physiological state of drivers, we used the skin potential signal, represented as S. Skin potential is one of the basic methods for the measurement of Electrodermal Activity (EDA), which is a term used to describe changes in the skin's ability to conduct electricity in response to, for example, stress or anxiety [2]. S was first down-sampled to 10 Hz and low-pass filtered using a second-order Savitzky-Golay smoothing filter with a length of 40.1 seconds, forming a smoothed skin potential G. Filter characteristics satisfactorily removed high-frequency noise from the raw signal. The G signal was then normalized as follows:

$$\tilde{G} = \frac{G - \mu_G}{\max(G)}. \quad (2)$$

μ_G and $\max(G)$ represent, respectively, the mean and maximum of skin potential of all training and testing data for a given driver. Let $\tilde{G}(n)$ represent the value of the n^{th} sample of the smoothed signal, where $n = 1, ..., N$, with N = 128. We investigated the following two statistical features:

1) Mean of normalized signal (mean skin potential):

$$f_1 = \frac{1}{N} \sum_{n=1}^{N} \tilde{G}(n) \quad (3)$$

2) Absolute value of the first-order difference of the normalized signal (Δ skin potential):

$$f_2 = \sum_{n=1}^{N} \left| \tilde{G}(n+1) - \tilde{G}(n) \right|. \quad (4)$$

Features were calculated for a window size N of 128 and shift of 5 points. f_1 was further quantized into four levels and f_2 into two.

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," which is defined as the inverse Fourier transform of the short-term log-power spectrum. Cepstral coefficients proved to be very effective in driver modeling and identification using pedal operation [8]. The cepstral coefficients were obtained as follows:

$$c(m) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{2\pi k m j / N}, \quad (5)$$
$$m = 0, 1, ..., N-1.$$

where $X(k)$ denotes the N-point discrete Fourier transform of the windowed signal $x(n)$.

Cepstral analysis is a source-filter separation process commonly utilized in speech processing. By keeping only the first several coefficients in the lower "quefrency" and setting others to zero, we can obtain a spectral envelope, the filter, which represents the process of acceleration or braking. On the other hand, a fine structure of the spectrum, the source, which works as the command signal for hitting a pedal, can be obtained by maintaining a higher "quefrency" range and setting the lower "quefrency" coefficients to zero. Cepstral coefficients $c(0)$ to $c(7)$ and their time derivative, calculated from a window of 1.0 seconds for brake and gas pedal signals, were utilized as features. Irritation level was linearly quantized into five levels; therefore driving behavior node was represented by five 32-dimensional Gaussians, one for each level. Before cepstral analysis, gas and brake signals were down-sampled to 10 Hz, and a median filter of 500 ms was utilized to remove spikes. Features were calculated using an analysis window of length $N = 128$ and shift of 5 points.

Transcription labels were also processed before being input in the network. To better model the influence of labels on drivers, anticipatory effects due to planning and persistence effects due to slow recovery should be taken into account. Using one of the basic operators in the field of mathematical morphology, called dilation, signals were dilated, i.e., had boundaries enlarged using a structure element of five seconds. The level of irritation, after quantization, was also dilated using the same structure element.

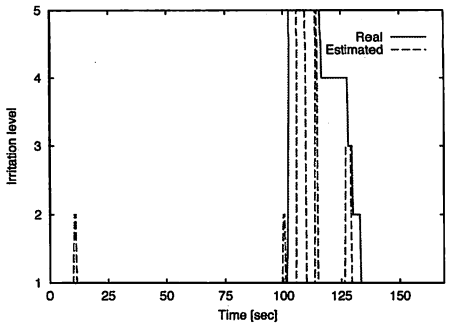

Fig. 3. Result of irritation estimation for driver 1.

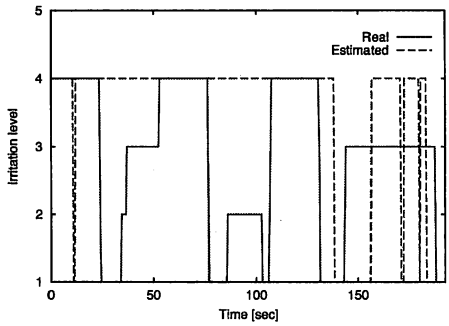

Fig. 4. Result of irritation estimation for driver 2.

D. Evaluation

We evaluated the capacity of the proposed system to detect irritation. Estimation and actual irritation, originally having five states, were quantized into two levels: irritated and not irritated. We then filtered the estimation using a median filter of ten seconds in order to reduce spikes and short gaps. In order to estimate the overall detection effectiveness, we added together true/false positives/negatives from all drivers, so that we could calculate a single point in the ROC space. We also specifically investigated the importance of the two proposed features: driving behavior and speech recognition, by training networks with and without information on these two nodes.

III. Results and Conclusions

Figures 3 and 4 show examples of estimation results. Solid lines indicate the actual level of irritation, while dashed lines indicate the estimate. Irritation observed from approximately 100 to 130 seconds was correctly detected in Fig. 3. Results shown in Fig. 4 are also satisfactory, although the classifier was unable to follow the actual level in great details. Networks trained without information on neither speech recognition nor driving behavior nodes achieved an overall true positive (TP) rate of 0.65 for an overall false positive (FP) rate of 0.19. When adding information on driving behavior and speech recognition, TP rate and FP rate were 0.71 and 0.13, respectively. Adding the new evidences improved the estimation. Results are reasonable and very encouraging, especially given the multiple challenges of a driver's affective state estimation under real driving conditions.

In this work we proposed a new approach to affective state

estimation. Our main contributions were: the design of a transcription protocol for various detailed events; the introduction of a new mathematical model for representing mutual dependencies between driver and driving context; and the investigation of the effect of pedal operation and speech recognition on irritation estimation—adding new evidences improved the estimation. Nevertheless, this is a pilot study and additional feature extraction methods have to be investigated, as well as the extension to dynamical BN.

REFERENCES

[1] G. M. Björklund. Driver irritation and aggressive behaviour. *Accident Analysis & Prevention*, 40(3):1069–1077, 2008.

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

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[4] M. Grimm et al. On the necessity and feasibility of detecting a driver's emotional state while driving. *Affective Computing and Intelligent Interaction*, pages 126–138, 2007.

[5] F. V. Jensen. *Bayesian networks and decision graphs*. Springer, 2001.

[6] C. Katsis et al. Toward emotion recognition in car-racing drivers: A biosignal processing approach. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 38(3):502–512, May 2008.

[7] J. McCall, S. Mallick, and M. Trivedi. Real-time driver affect analysis and tele-viewing system. *IEEE Intelligent Vehicles Symposium*, pages 372–377, 2003.

[8] C. Miyajima et al. Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, 95(2):427–437, 2007.

[9] K. P. Murphy. Inference and learning in hybrid Bayesian networks. Technical Report CSD-98-990, University of California, 1998.

[10] A. Ozaki et al. In-car speech data collection along with various multimodal signals. *Language Resources and Evaluation Conference (LREC)*, P22-9, May 2008.

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