Computers in Human Behavior 68 (2017) 450–455

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth

Enkelejda Kasneci [a,*], Thomas Kübler [a], Klaus Broelemann [b], Gjergji Kasneci [b]

[a] University of Tübingen, Department of Computer Science, Perception Engineering Group, Sand 14, 72076, Tübingen, Germany
[b] Schufa Innovation Lab, Kormoranweg 5, 65201, Wiesbaden, Germany

ARTICLE INFO

Article history:
Received 25 July 2016
Received in revised form
28 November 2016
Accepted 30 November 2016
Available online 2 December 2016

Keywords:
Eye tracking
Hazard perception
Physiological signals
Response aggregation
Eye tracking
Pupil dilation
ECG
Galvanic skin response

ABSTRACT

Today's driving assistance systems build on numerous sensors to provide assistance for specific tasks. In order to not patronize the driver, intensity and timing of critical responses by such systems is determined based on parameters derived from vehicle dynamics and scene recognition. However, to date, information on object perception by the driver is not considered by such systems. With advances in eye-tracking technology, a powerful tool to assess the driver's visual perception has become available, which, in many studies, has been integrated with physiological signals, i.e., galvanic skin response and EEG, for reliable prediction of object perception.

We address the problem of aggregating binary signals from physiological sensors and eye tracking to predict a driver's visual perception of scene hazards. In the absence of ground truth, it is crucial to use an aggregation scheme that estimates the reliability of each signal source and thus reliably aggregates signals to predict whether an object has been perceived. To this end, we apply state-of-the-art methods for response aggregation on data obtained from simulated driving sessions with 30 subjects. Our results show that a probabilistic aggregation scheme on top of an Expectation-Maximization-based estimation of source reliabilities can predict hazard perception at a recall and precision of 96% in real-time.

1. Introduction

Eye movements and physiological signals such as heart rate and galvanic skin conductance are measured in a variety of use-cases to derive information about a subject. More specifically, since the latter two signals are considered as strong indicators of cognitive load and stress, they have been analyzed in several applications to understand user behavior in complex tasks, and in particular during driving. In fact, sudden stress in safety-critical situations, as they may occur during driving, arouses the sympathetic nervous system. The subject transpires and skin conductance and heart rate change as a result (Backs, Lenneman, Wetzel, & Green, 2003; Helander, 1978; Lenneman & Backs, 2009; Lewis & Phillips, 2012; Mehler, Reimer, & Coughlin, 2012; Reimer, Mehler, Coughlin, Godfrey, & Tan, 2009; Son et al., 2011; Taylor, 1964). With advances in eye-tracking technology and analytical approaches, additional means have become available to assess performance and stress level during driving. More specifically, since changes in pupil diameter have been considered as a measure of emotional arousal and autonomic activation, pupil analysis has been employed in several studies (Benedetto et al., 2011; Bradley, Miccoli, Escrig, & Lang, 2008; Potamitis et al., 2000). The general assumption during driving is that visual perception of scene features such as signs, pedestrians, and obstacles requires foveated vision, i.e., an object of interest in the scene is considered as perceived if it has been fixated by the driver (Fletcher & Zelinsky, 2009). Although peripheral vision is considered as sufficient for some subtasks, such as keeping the vehicle centered in the lane (Summala, Nieminen, & Punto, 1996), it has been reported that peripheral vision is insufficient for the detection of traffic hazards (Maltz & Shinar, 2004).

Recent studies investigating the correlation between hazard fixation and its perception have however reported that the direction of a driver's gaze towards an upcoming hazard does not a priori imply its perception (Kasneci, Kasneci, Kübler, & Rosenstiel, 2015, pp. 411–434; Tafaj, Kübler, Kasneci, Rosenstiel, & Bogdan, 2013). Moreover, several studies have reported that subjects have shown adequate hazard detection although the target object has not been fixated (Kasneci et al., 2014; Kübler et al, 2015a, b). Thus, in some cases, peripheral vision may be sufficient for hazard perception

* Corresponding author.
E-mail address: enkelejda.kasneci@uni-tuebingen.de (E. Kasneci).

http://dx.doi.org/10.1016/j.chb.2016.11.067
0747-5632/

(Tafaj et al., 2013).

In driving scenarios, determining whether a hazard was perceived by the driver early on can lead to significant support of automatic accident avoidance systems. To approach this challenge, signals from different sensors need to be processed in an online fashion and aggregated according to their reliability. The reliability of a sensor, however, depends not only on the type of the sensor but also on the subject. For example, eye-tracking signal is sensitive to make-up and changing illumination, while skin conductance and heart rate are vulnerable to loosened or detached electrodes. Deriving a binary decision about hazard perception from raw sensor data is a challenging problem of its own and requires device specific filtering, synchronization and processing. We will threat this necessary preprocessing step as abstract throughout the manuscript and work only on the readily preprocessed binary decision label in order to demonstrate the proposed concepts in a more general, device independent way. For details on how the data used throughout the manuscript was preprocessed, please refer to (Kübler et al., 2014).

In this paper we address the problem of how to combine multiple physiological (i.e., heart rate and galvanic skin response) signals with eye-tracking measurements for an automated detection of target perception. Especially in the absence of ground truth, a viable aggregation method has to jointly infer the reliability of the source delivering the signal and the true event taking place (e.g., whether the target was perceived) (Fig. 1).

From a theoretical viewpoint the problem can be formalized as follows. Given binary signals or responses $X_{i1}, \ldots, X_{in} \in \{0, 1\}$ from n sources (e.g., sensors), 1 meaning that the event $e_i, 0 \le i \le t$, has taken place and 0 that e_i has not occurred, how can we aggregate the responses in a way that we can learn the latent truth (i.e., whether the event occurred or not).

In the data mining literature, there is a vast body of work addressing the aggregation of responses in order to find the latent ground truth. Some of these approaches can be adapted to the aggregation of sensor signals. However, only a few are applicable in real-time. In this paper we analyze the most popular algorithms in this realm and provide a practical guidance for their real-time application in driving scenarios.

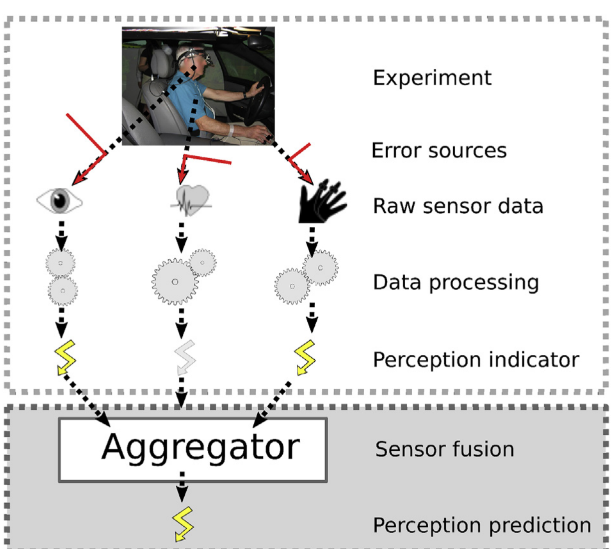


Fig. 1. This manuscript discusses the aggregation of physiological sensor data (eye-tracking, heart rate, skin conductance) that has already been preprocessed to a binary perception indicator (displayed as a thunderbolt). Aggregation and reliability estimation of the individual perception indicators is performed (bottom box). For information on recording and preprocessing of the data (top box) see, for example (Kübler et al., 2014).

In the following we will use the terms signal, response, answer and claimed value interchangeably.

The rest of the paper is organized as follows: Section 2 gives an overview of related work in the area of latent truth discovery and reliable response aggregation. Section 3 provides a practical framework for the real-time application of popular truth discovery and aggregation algorithms. An extensive analysis and evaluation of the algorithms on real-world data collected from driving experiments with human subjects is presented in Section 5. The data was carefully labeled by experienced annotators as described in Section 4.

2. Related work

From an abstract viewpoint, there are 3 categories of latent truth discovery methods:

Bayesian Inference algorithms use prior distributions for the truth and reliability parameters and jointly estimate truth and source reliability by fitting the parameters to the available data based on the assumed prior distributions.

Fix-point and Expectation Maximization algorithms start with an initial guess on the truth and reliability parameters and simplifying assumptions are used to iteratively fit the parameters to the available data.

Semi-Supervised Learning algorithms start with a set of known ground truth labels. This initial ground truth and other assumptions are exploited to learn the reliability of sources. In turn, the reliability estimations can be used to estimate the latent truth.

In the following paragraphs, we give an overview of the above three groups by highlighting representative approaches.

2.1. Bayesian Inference

TruthFinder (Yin, Han, & Philip, 2008) models the influence between claimed values and applies Bayesian analysis to iteratively estimate source reliabilities and the latent truth. AccuSim (Dong, Berti-Equille, & Srivastava, 2009; Li, Dong, Lyons, Meng, & Srivastava, 2012) integrates the similarity between claimed values into the Bayesian inference approach and proposes an extension of the algorithm AccuCopy in which also source similarities – in terms of which source might have copied from which other source – are considered. The more a source has copied from other sources, the more its weight is reduced.

A Bayesian approach to knowledge corroboration is proposed by Kasneci, Van Gael, Herbrich, & Graepel (2010); Kasneci, Van Gael, Stern, Graepel (2011), where a latent truth discovery model integrates the logical dependencies between facts in a knowledge base and crowd opinions to derive the underlying correctness of the facts in the knowledge base.

Latent Truth Model (LTM) (Zhao, Rubinstein, Gemmell, & Han, 2012) is a probabilistic graphical model that applies collapsed Gibbs sampling to estimate the false positive and the false negative rate of sources by optimizing for the most probable answers.

Another Bayesian inference approach for continuous responses is presented in Zhao & Han, 2012.

2.2. Fix-point algorithms and expectation maximization

In 2-Estimates (Galland, Abiteboul, Marian, & Senellart, 2010) the assumption that there is one and only one true value for each object is integrated in a voting-based fix-point algorithm. The authors also propose an extension, 3-Estimates, in which the difficulty of deriving the true value of an object is considered.

In Pasternack & Roth (2010), a source uniformly "invests" its reliability among the values it has claimed for the objects. The confidence of a value grows according to a non-linear function defined on the sum of invested reliabilities from the sources that claimed it. In turn, the sources collect credits back from the confidence of their claimed values.

A Maximum Likelihood formulation of latent truth discovery for crowd/social sensing applications is provided by Wang, Kaplan, Le, & Abdelzaher (2012). The Expectation Maximization (EM) algorithm is proposed to derive the most probable answers as well as the true positive and true negative rates of sources (human agents in this case).

2.3. Semi-supervised learning

In Yin & Tan (2011) a semi-supervised truth discovery approach is proposed. An initial set of known ground truth labels is used to estimate the reliability of sources. The formalization of mutual exclusivity and mutual support between claimed values are exploited to capture the relations between values and to guide the algorithm towards reliability and truth estimations.

3. Methods

In this section, we present the approaches that we have evaluated in the driving context for the aggregation of physiological and eye tracking signals. Alongside a majority voting aggregation scheme, we have selected Latent Truth Model (LTM) (Zhao et al., 2012) as a representative of Bayesian Inference approaches and the approach of (Wang et al., 2012) as a representative of fix-point and Expectation Maximization algorithms. Additionally, we provide an optimal aggregation scheme that can be applied in real-time when true positive rates and true negative rates are available.

3.1. Majority voting (MV)

In the absence of ground truth, majority voting is the simplest way to aggregate the responses of independent expert sources. Given the responses $X_1, \ldots, X_n \in \{1, \ldots, K\}$ of n independent expert sources (e.g., well-calibrated sensors) for an event e, the truth is derived as $\max_{k\in\{1,\ldots,K\}} |\{j : X_j = k\}|$. Note that we omit the event-index i for the sake of readability. Despite its simplicity, this method produces satisfactory results in many practical cases. Because of its simplicity, the method is highly efficient and applicable to real-time scenarios.

3.2. Expectation maximization (EM)

A formulation of the truth discovery task as maximum likelihood estimation problem is proposed by the authors of Wang et al (2012). The solution to the problem is given by a regular expectation maximization algorithm which iteratively estimates the reliability of information sources (in terms of their true positive and true negative rates) and the truth of facts until convergence.

3.3. Latent Truth Model (LTM)

LTM (Zhao et al., 2012) is a probabilistic graphical model which estimates two types of errors for each source, under the assumption of multiple truths, in a Bayesian manor: the false positive rate and the false negative rate. These rates are in turn used to estimate the truth of facts for which the information sources may provide conflicting claims (i.e., fact is true or fact is false).

Unfortunately, both EM and LTM are computationally intensive, which makes their real-time application in the driving context difficult. However, both methods can be used in the background to estimate source reliability parameters such as true positive and true negative rates. The better these rates are estimated, the more accurate are the results obtained by the following real-time aggregation scheme on top of the estimated true positive and true negative rates.

3.4. An optimal aggregation scheme (OAS)

If we would know the true reliability of each signal source – in terms of its true positive rate and true negative rate – we could aggregate the source responses in a probabilistically optimal way as follows:

Assume each signal source s_j provides a response $X_j \in \{1, \ldots, K\}$. Given the responses of n sources on the occurrence of a discrete event $e \in \{1, \ldots, k\}$, the most probable event to have occurred would be

$$\arg\max_k p(e = k | X_1, \ldots., X_n) = \arg\max_k \frac{p(X_1, \ldots, X_n | e = k) p(e = k)}{p(X_1, \ldots, X_n)}$$

Assuming independence between the signals given the event, we have

$$\arg\max_k p(e = k | X_1, \ldots., X_n) = \arg\max_k \frac{\prod_j p(X_j | e = k) p(e = k)}{p(X_1, \ldots, X_n)}$$

$$\propto \arg\max_k \left(\prod_{\substack{j=1,\ldots,n \\ X_j = k}} p(X_j = k | e = k) \prod_{\substack{j=1,\ldots,n \\ X_j \neq k}} p(X_j \neq k | e = k) \right) p(e = k)$$

Note that since $p(X_1, \ldots, X_n)$ is independent of k it does not influence the choice of the most probable k. 0 Following the above derivation, in the case of binary events and responses from {0,1}, we have

$$p(e = 1 | X_1, \ldots., X_n) \propto \left(\prod_{\substack{j=1,\ldots,n \\ X_j = 1}} tpr_j \prod_{\substack{j=1,\ldots,n \\ X_j = 0}} fnr_j \right) p(e = 1)$$

and analogously

$$p(e = 0 | X_1, \ldots., X_n) \propto \left(\prod_{\substack{j=1,\ldots,n \\ X_j = 0}} tnr_j \prod_{\substack{j=1,\ldots,n \\ X_j = 1}} fpr_j \right) p(e = 0)$$

The above aggregation follows the Naive Bayes principle, and, if the sensors' responses are indeed independent given the event, it is optimal for the choice of the most probable event in a probabilistic sense. In practice, in the presence of multiple sensors, a sum of logarithms would be used to avoid computational arithmetic underflow.

The presented aggregation scheme can be used on top of complex and computationally intensive algorithms, such as the Expectation Maximization approach – which, in turn, can be run in the background to produce estimations of the true positive and true negative rates from sampled data. The real-time aggregation of responses is provided by the above scheme.

4. Experimental data

4.1. Experimental setting

Thirty subjects were enrolled in a driving simulator study that was conducted in the moving base driving simulator at the Mercedes-Benz Technology Center in Sindelfingen, Germany (Fig. 2(a)). The cabin contained a real car body (Mercedes S class with automatic transmission) amidst a 360° projected virtual reality. Acceleration, sound effects and car environment contributed to a near-to-reality driving experience.

Each subject completed a 40 min drive (37.5 km length) in the simulator. The driving route contained rural as well as urban areas with different speed limits up to 100 km/h. Nine hazardous traffic situations, e.g., pedestrians suddenly appearing behind parking vehicles and trying to cross the road or risky overtaking maneuvers, were placed along the driving course. Fig. 2(b) shows an example of such a hazardous situation that arises as a pedestrian tries to cross the road from the left. In this case, if the driver does not react to the appearance of the pedestrian, the simulation avoids the crash, e.g., by having the pedestrian quickly leap backwards. Similar crash avoidance strategies are simulated for all other hazards; hence, (additional) psychological stress of the driver can be mitigated. A detailed description of this study was previously published (Kübler et al., 2014). Eye movements were recorded by means of a mobile, head-mounted Dikablis eye tracker (Ergoneers GmbH) at 25 frames per second simultaneously to the physiological parameters galvanic skin conductance (GSC) and heart rate (ECG). Skin conductance was recorded via electrodes at the fingertips using a Biotrace+ system. ECG was recorded by a mobile 3-channel device of type custo med EKG.

4.2. Physiological data

For each hazardous situation, the driver's reaction to a traffic hazard was rated as passed or failed by a driving instructor. If the driver did not show an adequate reaction to the approaching hazard (e.g., by braking, or by taking an appropriate collision avoidance action), the specific situation was rated as *failed*, otherwise as *passed*. The analysis of the physiological and eye-tracking signal will be described briefly in the following. A detailed description of the methods employed for this analysis can be found in (Kübler et al., 2014). Aim of this work is to find an optimal aggregation method for the sensor responses according to their reliability to infer whether a hazard was perceived by the driver.

4.2.1. ECG analysis

In order to detect stress-related increases in heart rate, large and fast increments in heart rate were extracted. For subjects with a higher overall variation in heart rate, a higher threshold is required in order to classify an increase as unusually large. The threshold was set to three standard deviations of the individual's heart rate (for an assumption of normal distribution, over 99% of the data should be contained within this interval). Thus, an increase of more than three standard deviations was considered as a stress response to a hazardous situation.

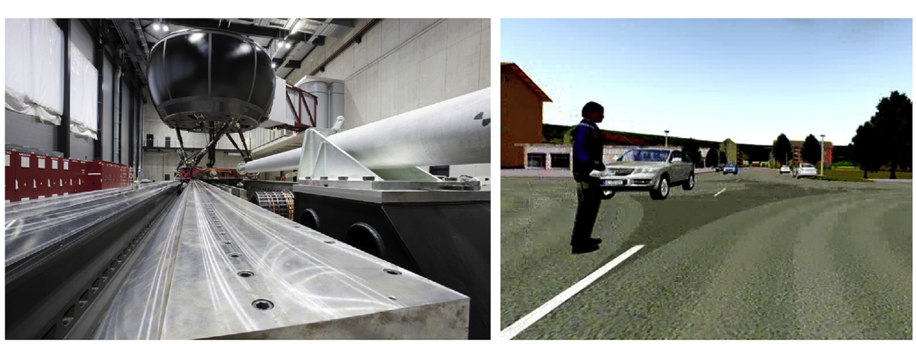
4.2.2. GSC analysis

GSC data was first smoothed by a Butterworth low-pass filter. A threshold was then applied to the change in conductance, similar to the procedure introduced by Healey, Seger, & Picard (1999). By investigating only the change between small time intervals, effects such as a steady increase in absolute value over the whole experiment are equally distributed over the whole recording and become negligible. Similar to the extraction of heart rate increments, the threshold was set to three standard deviations of GSC change.

4.2.3. Alignment of the driving scene with the sensor data

The spatial extent and the position of the traffic hazards in the scene was manually annotated using bounding boxes, i.e., a rectangular border enclosing the hazardous object in the scene image. For this purpose, the hazardous situations were analyzed starting from the moment when they entered the driving scene up to when the hazardous situation was resolved (e.g., the driver reacted to the approaching hazard or the simulation resolved the hazardous situation). The annotation was performed manually and frame-wise on the video recordings of the driving scene.

The fixations of the driver were extracted from the raw eye-tracking data in an online fashion based on an Bayesian mixture model (Kasneci, Kasneci, Kübler & Rosenstiel, 2014; Tafaj, Kasneci, Rosenstiel, & Bogdan, 2012). The spatial extent of a driver's fixation was then approximated by an ellipse, thus marking the object of interest viewed by the driver. Finally, the resulting bounding boxes (around hazardous objects) were matched with the fixation ellipses (that describe the focus of the driver). As in Kasneci et al. (2015), pp. 411–434, whenever a fixation of the driver intersected



(a) (b)

Fig. 2. Moving base driving simulator (Zeeb, 2010, pp. 157–165). (a) the entire cabin is mounted on a hexapod, moving along a 12*m* rail (up to 1g acceleration force). (b) Virtual reality scene as seen from inside the cabin. A pedestrian intends to cross the road from the left side (Kübler et al., 2014). Figures provided by Daimler AG.

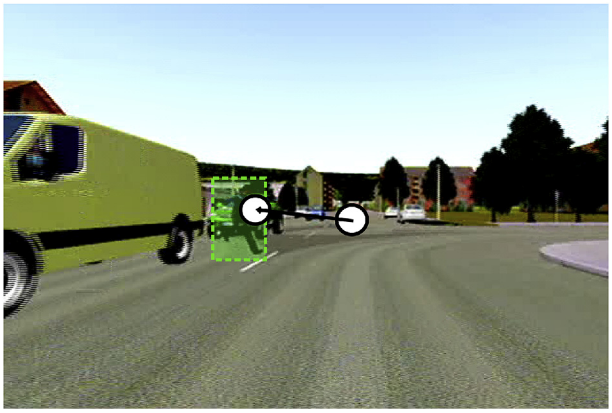


Fig. 3. Hazardous situation (pedestrian crossing the road between cars) and fixation data indicated as a sequence of white circles. Once the bounding box around the pedestrian was hit by a fixation, the hazard was registered as *seen*.

a bounding box, a traffic hazard was considered as seen by the driver, see Fig. 3

4.2.4. Pupil data

To extract pupil dilation due to stress from the pupil signal, we employed a Wavelet-based analysis. Relative energies of the Daubechies wavelet transformed signal were used in a machine learning step in order to distinguish between pupil size changes caused by the ambient illumination and due to stressful events. This SVM classification step results in a binary response of whether a stress peak was detected or not.

5. Results

In total, there were 270 hazardous situations, i.e., nine for each of the 30 subjects. Six subjects aborted the driving session early due to motion sickness or technical problems. In some cases, skin conductance, ECG, or pupil signal could not be analyzed due to missing data.

Table 1 presents a detailed description of the signal responses. The drivers showed an adequate driving response in 227 out of 237 hazardous situations. A failure to react to the approaching hazard occurred in only 10 situations.

As summarized in Table 1, the galvanic skin response was present (i.e., a change was detected in the sensor signal, encoded as a 1) in 193 out of 227 situations where the driver reacted adequately (*passed*). With regard to the class *passed*, these responses represent the true positives (TP). In 16 *passed* situations, no change in GSC was detected. These cases represent false negatives (FN). In the remaining 18 *passed* situations, a GSC signal was not available. For the situations that were rated as *failed*, no change in the GSC signal was detected in 6 out 10 hazardous situations (true negatives, TN). In the remaining 4 cases, a response of the GSC signal was registered, although the driver did not react (false positives, FP). The responses of the ECG signal, pupil dilation, and hazard fixation are given in Table 1 analogously.

Table 1
Sensor responses in 227 *passed* and 10 *failed* situations.

	Passed (n = 227)			Failed (n = 10)		
	TP	FN	Non-available	TN	FP	Non-available
GSC change	193	16	18	6	4	0
ECG change	164	55	8	7	3	0
Hazard fixated	211	9	7	3	7	0
Pupil dilation	190	21	16	1	8	1

To compute the reliability of the sensor signals and to predict hazard perception based on the data summarized in Table 1, we employed the algorithms *Majority Voting*, *LTM*, and *EM* as introduces previously.

Table 2 presents the source reliabilities as computed by the different algorithms. The results are given in terms of the true positive rate (tpr) and false positive rate (fpr). According to these results, the most reliable physiological signal is the fixation-based analysis (Hazard fixated), with true positive rates varying between 87.32% (Majority voting) and 99.58% (LTM).

Table 2
Source reliabilities as computed by the different algorithms.

	Majority Voting [%]		LTM [%]		EM [%]	
	tpr	fpr	tpr	fpr	tpr	fpr
GSC Change	86.07	13.93	96.5	25.05	93.5	18.96
ECG Change	78.6	21.4	75.4	63.99	74.8	63.29
Hazard fixated	87.32	12.68	99.58	20.31	96.72	12.17
Pupil dilation	84.09	15.91	89.81	39.13	87.8	34.34

Table 3 presents the predictability of hazard perception for each of the three employed techniques. All of them learn not only the reliability of each source, but also deal with non-availability of the sources. Among the applied aggregation techniques, LTM and EM show similar precision and recall values with regard to hazard perception. The last two approaches in Table 3 OAS_LTM and OAS_EM correspond to Optimal Aggregation Scheme (Section 3.4) that were applied on top of the LTM and EM method, respectively.

Table 3
Predictability. *If signal available.

Method	Precision [%]	Recall [%]
Hazard fixated*	95.47	95.90
Majority Voting	91.36	96.52
LTM	94.78	94.78
EM	94.78	94.78
OAS_LTM	96.84	94.71
OAS_EM	96.05	96.47

6. Discussion

In case of fully reliable sensor signals, we would expect a response from each source upon hazard perception (i.e., in the case of passed situations). More specifically, since we assume that hazard perception induces stress, we would expect to measure a change in GSC, ECG, and a dilation of the pupil. Furthermore, following the assumption of foveated vision, we would assume an intersection of the fixation location with the hazardous object. Analogously, in case of no driver reaction, we would expect no response in any of the signals. However, Table 1 shows that the signal sources are not reliable. Furthermore, te availability of signals is not always guaranteed. Hazard fixation, the most predictive indicator, is available in 230 of 237 situations. But for GSC change and pupil dilation the availability is much worse.

Despite the very high true positive rate, LTM might not be the best choice among this three techniques, since it has higher false positive rates than the other two approaches.

In difference to the EM algorithm, our aggregation scheme estimates the prevalence of 1 (or 0) simply by the relative count of responses equal to 1 (or 0). In addition, our aggregation scheme can be applied also to cases where none of the sources is available. By applying such an aggregation scheme, the result can be improved by a margin of 2% in comparison to LTM and EM. In summary, OAS_EM represents the best aggregation approach for the problem and data at hand.

The pupil dilation did not predict hazard perception reliably. As it is influenced by a variety of other factors, such as illumination and cognitive load, natural environments make it very difficult to extract the pupil dilation component caused by stress.

According to our results, the fixation-based analysis (Hazard fixated) represents the most reliable signal source among the available physiological signals. In fact, for the situations in which this signal is available, it can predict hazard perception with a precision of 95.47% and recall of 95.9%, Table 3.

In contrast to the eye-tracking signal which has a negligibly small delay in the order of few ms, the detection of changes in the vital parameters ECG and GSR is only possible with a relatively long delay (approximately $1-4$ seconds). Thus, these signals can be used as input and predictors for triggering driver assistance systems only to a limited extent.

7. Conclusion

Automated, real-time aggregation of sensor measurements is important for the task of event recognition in many dynamic scenarios. For the recognition of hazard perception in the driving context, this paper provides a strong first basis for the real-time aggregation of physiological and eye-tracking signals – the latter delivering reliable indicators of hazard perception.

Beyond today's driving assistance, our findings can be beneficial in the context of autonomous driving to determine the situation awareness and attentiveness of the driver.

References

Backs, R. W., Lenneman, J. K., Wetzel, J., & Green, P. (2003). Cardiac measures of driver workload during simulated driving with and without visual occlusion. *Human Factors, 45*(4), 525–538.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research F: Traffic Psychology and Behaviour, 14*(3), 199–208.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602–607.

Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment, 2*(1), 550–561.

Fletcher, L., & Zelinsky, A. (2009). Driver inattention detection based on eye gaze-road event correlation. *The International Journal of Robotics Research, 28*(6), 774–801.

Galland, A., Abiteboul, S., Marian, A., & Senellart, P. (2010). Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining, ACM* (pp. 131–140).

Healey, J., Seger, J., & Picard, R. (1999). Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. *Biomedical Sciences Instrumentation, 35*, 193–198.

Helander, M. (1978). Applicability of drivers' electrodermal response to the design of the traffic environment. *Journal of Applied Psychology, 63*(4), 481.

Kasneci, G., Van Gael, J., Herbrich, R., & Graepel, T. (2010). Bayesian knowledge corroboration with logical rules and user feedback. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 1–18). Springer.

Kasneci, G., Van Gael, J., Stern, D., & Graepel, T. (2011). Cobayes: Bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Proceedings of the fourth ACM international conference on Web search and data mining, ACM* (pp. 465–474).

Kasneci, E., Kasneci, G., Kübler, T., & Rosenstiel, W. (2014). The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes. In *Proceedings of the symposium on eye tracking research and applications* (pp. 323–326). ACM.

Kasneci, E., Sippel, K., Aehling, K., Heister, M., Rosenstiel, W., Schiefer, U., et al. (2014b). Driving with binocular visual field Loss? A study on a supervised on-road parcours with simultaneous eye and head tracking. *Plos One, 9*(2), e87470.

Kasneci, E., Kasneci, G., Kübler, T. C., & Rosenstiel, W. (2015). Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception. In *Artificial Neural Networks* (pp. 411–434). Springer International Publishing.

Kübler, T. C., Kasneci, E., Rosenstiel, W., Schiefer, U., Nagel, K., & Papageorgiou, E. (2014). Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss. *Transportation Research Part F: Traffic Psychology and Behaviour, 24*, 231–243.

Kübler, T. C., Kasneci, E., Rosenstiel, W., Heister, M., Aehling, K., Nagel, K., et al. (2015a). Driving with glaucoma: Task performance and gaze movements. *Optometry & Vision Science, 92*(11), 1037–1046.

Kübler, T. C., Kasneci, E., Rosenstiel, W., Aehling, K., Heister, M., Nagel, K., et al. (2015b). Driving with homonymous visual field defects: Driving performance and compensatory gaze movements. *Journal of Eye Movement Research, 8*(5).

Lenneman, J. K., & Backs, R. W. (2009). Cardiac autonomic control during simulated driving with a concurrent verbal working memory task. *Human Factors, 51*(3), 404–418.

Lewis, M., & Phillips, J. (2012). Older people's cardiac responses as indicators of stress in familiar and unfamiliar environments. *Psychophysiology, 49*(4), 478–483.

Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2012). Truth finding on the deep web: Is the problem solved? *Proceedings of the VLDB Endowment, 6*, 97–108.

Maltz, M., & Shinar, D. (2004). Imperfect in-vehicle collision avoidance warning systems can aid drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(2), 357–366.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(3), 396–412. http://dx.doi.org/10.1177/0018720812442086.

Pasternack, J., & Roth, D. (2010). Knowing what to believe (when you already know something). In *Proceedings of the 23rd international conference on computational linguistics* (pp. 877–885). Association for Computational Linguistics.

Potamitis, T., Slade, S. V., Fitt, A. W., McLaughlin, J., Mallen, E., Auld, R. J., et al. (2000). The effect of pupil dilation with tropicamide on vision and driving simulator performance. *Eye, 14*, 302–306.

Reimer, B., Mehler, B., Coughlin, J., Godfrey, K., & Tan, C. (2009). An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications, ACM* (pp. 115–118).

Son, J., Mehler, B., Lee, T., Park, Y., Coughlin, J., & Reimer, B. (2011). Impact of cognitive workload on physiological arousal and performance in younger and older drivers. In *Proceedings of the sixth international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 87–94).

Summala, H., Nieminen, T., & Punto, M. (1996). Maintaining lane position with peripheral vision during in-vehicle tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 38*(3), 442–451.

Tafaj, E., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2012). Bayesian online clustering of eye movement data. In *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12, ACM* (pp. 285–288). http://dx.doi.org/10.1145/2168556.2168617.

Tafaj, E., Kübler, T., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2013). Online classification of eye tracking data for automated analysis of traffic hazard perception. In *Artificial Neural Networks and Machine Learning ICANN 2013* (Vol. 8131, pp. 442–450). Berlin Heidelberg: Springer.

Taylor, D. (1964). Drivers' galvanic skin response and the risk of accident. *Ergonomics, 7*, 439–451.

Wang, D., Kaplan, L., Le, H., & Abdelzaher, T. (2012). On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on information processing in sensor networks, ACM* (pp. 233–244).

Yin, X., & Tan, W. (2011). Semi-supervised truth discovery. In *Proceedings of the 20th international conference on world wide web, ACM* (pp. 217–226).

Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *Transportation Research Part F: Traffic Psychology and Behaviour, 20*(6), 796–808.

Zeeb, E. (2010). *Actes INRETS*. Daimlers new full-scale, high-dynamic driving simulator–a technical overview.

Zhao, B., & Han, J. (2012). A probabilistic model for estimating real-valued truth from conflicting sources. In *Proceeding of QDB*.

Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment, 5*(6), 550–561.