Action Units and Their Cross-Correlations for Prediction of Cognitive Load during Driving

Anıl Yüce, *Student Member, IEEE,* Hua Gao, Gabriel L. Cuendet, *Student Member, IEEE,* and Jean-Philippe Thiran, *Senior Member, IEEE*

Abstract—Driving requires the constant coordination of many body systems and full attention of the person. Cognitive distraction (subsidiary mental load) of the driver is an important factor that decreases attention and responsiveness, which may result in human error and accidents. In this paper, we present a study of facial expressions of such mental diversion of attention. First, we introduce a multi-camera database of 46 people recorded while driving a simulator in two conditions, baseline and induced cognitive load using a secondary task. Then, we present an automatic system to differentiate between the two conditions, where we use features extracted from Facial Action Unit (AU) values and their cross-correlations in order to exploit recurring synchronization and causality patterns. Both the recording and detection system are suitable for integration in a vehicle and a real-world application, e.g. an early warning system. We show that when the system is trained individually on each subject we achieve a mean accuracy and F-score of $\sim 95\%$, and for the subject independent tests $\sim 68\%$ accuracy and $\sim 66\%$ F-score, with person-specific normalization to handle subject dependency. Based on the results, we discuss the universality of the facial expressions of such states and possible real-world uses of the system.

Index Terms—Affect sensing and analysis, Facial expression, Affective computing applications, Vehicle operation, Emotional corpora, Driver cognitive distraction.

1 Introduction

Driver monitoring in real-time is an emerging topic thanks to the availability of faster software and smaller hardware that can easily be integrated in consumer vehicles. In addition to systems that record and analyze driving data, e.g. wheel movement, speed and acceleration, or driver's physiological signals, research on visual monitoring systems are also on the rise and such systems will soon be more and more frequently integrated in automobiles on the market. In this work, we propose a visual driver monitoring system that aims to detect if the driver is under a secondary *cognitive load* using the facial behavior, and that is trained and tested on a driving simulator, but is suitable to be integrated in a vehicle to provide real-time information.

There has been a long discussion on how to define driver distraction. Pettitt et al. [1], Lee et al. [2] and more recently Regan et al. [3] have published works on how to define the term and compile the existing definitions. Lee et al. summarize it as the diversion of attention away from activities critical for safe driving toward a competing activity [2]. In [1] there is a more extensive definition and driver distraction is stated as a delay by the driver in the recognition of information necessary to safely maintain the driving task, due to some event, activity, object or person, within or outside the vehicle that compels or tends to induce the driver's shifting attention away from fundamental driving tasks by compromising the driver's auditory, biomechanical, cognitive or visual faculties, or combinations thereof ([3], [4]). Driver distraction can be in three types depending on its source and demand: visual, manual and cognitive [5]. Auditory distractions are sometimes considered as a fourth type of distraction, yet in this work we consider them also as cognitive distraction. Even though we introduce a database that includes all three types of distraction, our automatic system focuses on detecting the cognitive type, which can be defined as a diversion of the driver's attention from the driving task, not necessarily requiring any sharing of visual processing or involving or demanding a biomechanical action. It includes internally induced distraction, such as mind wandering or daydreaming but excludes cases such as boredom, sleepiness, or driving under the influence of alcohol or drugs and substances that alter the mental state. Throughout the rest of the article we use both the terms cognitive distraction and cognitive load, the latter referring to the state whose presence we aim to detect. It can be considered as a sub-group of what we have aimed to induce in the driving experiments, that is cognitive distraction.

Many studies show that driver distraction is one of the most important causes of traffic accidents, along with alcohol use and speeding. A study conducted in France showed that 17% of 453 accidents that resulted in admittance to the emergency room was caused by a high mental distraction of the responsible driver [6]. More recent studies from the same group has shown that induced distractive thoughts led to less micro-regulation of both speed and lateral position and narrowed visual scanning of the driving scene [7], that mind wandering is the cause of 8% of close to 1000 accidents according to emergency room interviews with the drivers [8] and that it affects 85.2% of the drivers especially in situations requiring less attention from the driver such as an everyday commute or a monotonous motorway [9].

In [10], the authors have collected and analyzed almost 43000 hours of driving data and shown that 78% of the crashes and 65% percent of near-crash incidents involve

- *The authors are with the Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland. E-mail: {anil.yuce@epfl.ch, hua.gao@epfl.ch, gabriel.cuendet@epfl.ch, jean-philippe.thiran@epfl.ch}*

driver inattention due to various secondary tasks. A similar study sponsored by the United States department of transportation showed that drivers investigated were engaged in non-driving related tasks in 71% of crashes [11]. Again in the U.S., it is estimated that around 20% of all police reported road crashes involve driver distraction as a contributing factor [12]. In [13] the authors provide a large scale examination of the relationship between driver distraction and driver errors, along with a list of existing studies.

Even though the numbers differ depending on the type and amount of data analyzed in each study, they all show that internally or externally caused driver distraction is a very critical risk factor. However, there is no clear distinction on which kind of distraction is more dangerous or happens more frequently. Indeed, usually the three types happen in combinations of two or three making this comparison even more difficult. We focus on the detection of cognitive distraction for two main reasons. Firstly, it provides a big challenge in terms of computer vision and machine learning as it does not have such clear indicators as visual or manual distraction. Secondly, cognitive distraction is a state with direct links to affective science and with complex repercussions in terms of facial expressions, which makes it even more so intriguing to investigate. We, therefore, address the problem of automatic detection of cognitive driver distraction using visual monitoring of the driver's face and propose a system that is tested on simulation data and that can easily be integrated in real cars for applications like an early alert system or activation of countermeasures in order to help the driver regain his attention on the driving task.

In this article, we present two main contributions. First, we introduce the EPV-DIST database, which is a multi-camera visual database of 46 people driving a simulator with different distraction conditions induced. The recording setup for our database has been planned to represent a configuration that is feasible to place in a car and work robustly in different light conditions. Second, we propose a method based on Facial Action Units (AU) to detect the conditions where the drivers were induced cognitive distraction. AUs model every unitary muscle movement on the face and were proposed as a means of defining and quantifying facial actions in an objective way [14]. The proposed system first generates a virtual frontal view from the three frames captured by the multi-camera system, then applies the AU detection we have previously proposed in [15]. Then, we extract features from the dynamic continuous value outputs of the AU detection system, independently for each 14 AU detected, and also from their cross-correlations at different delay points. This second type of features allows for exploiting the inter-relations of AUs and their synchronization behavior in different conditions, with the hypothesis that it will improve detection of facial expressions that cannot easily be defined as in the six basic expressions. All features are then fed-into Support Vector Machine (SVM) or Random Forest (RF) classifiers to obtain a decision on each sequence that has been labeled as being recorded under cognitive load or not.

The following sections of the paper are organized as follows: In Sec. 2 we give a review of existing algorithms and applications for visual driver monitoring. In Sec. 3 we introduce the EPV-DIST database and detail the recording setup and the experiment protocol. Sec. 4 describes all components of the cognitive load detection system, with the results and findings presented in Sec. 5. Finally, we conclude the paper in Sec. 6 with a discussion on the system capabilities and possible future improvements and applications.

2 Related Work

In this section we present a brief review of existing work on visual driver monitoring for various applications as well as different modalities used for detecting various types of driver distraction. This review is not focused on automatic distraction detection during driving because of the rather low number of such work in the literature and in order to provide the reader a general insight on similar automatic systems. An extensive review is also given in [16] and [17], the interested reader is referred to these publications for more approaches and applications not listed here.

Over the years most of the research on visual driver monitoring systems have focused on fatigue detection, which is another critical factor for human error in driving and can be considered related to cognitive distraction, yet excluded from the definition that we adopt for cognitive distraction and load (see Sec. 1). In [18] an automatic mouth movement analysis was performed to detect fatigue related actions, and also speaking, while in [19] AUs were used within a Dynamic Bayesian Network (DBN) to detect driver vigilance. The head pose dynamics have also been successfully exploited in a real-time driver awareness detection system [20]. Another commonly used visual cue for fatigue detection is the Percent Eye Closure Measure (PERCLOS), as used for instance in [21]. An approach on fatigue detection, rather close to ours is the work by Vural et al. [22], where the authors used many AUs, including head-pose, and analyzed their relation to fatigue during a three-hour simulator driving experiment after midnight. As expected, the most relevant features were related to eye-blink (AU45) and also outer brow raise (AU2), as the subjects tried to remain awake. This work is particularly interesting as we are able to compare the outcome of the relevant features analysis.

As for automatic detection of distraction, a non-vision based system is presented in [23] where the driving information, such as the speed, position of the pedal and steering wheel have been used to detect visual distraction tested with various machine learning methods. Wöllmer et al. also used the driving information and non-vision based head tracking data to detect cases of visual distraction while performing various tasks [24]. In [25] the authors used eye movements and driving performance data within a Bayesian Network framework. The system can detect $\sim 80\%$ of distraction cases that are due to interacting with an in-vehicle information system (IVIS). A similar study is presented in [26] where the gaze angle and fixation data was used to recognize distraction induced by the IVIS. In [27] the eye movements are analyzed to predict visual inattention using Neural Networks. The gaze information was used along with head movements and lane position of the vehicle in [28] to detect induced visual and cognitive distraction using a stereo-vision system integrated in trucks

(a) The three-camera acquisition setup

(b) Driver's position during the recording

Fig. 1. The recording setup of the experiment

and passenger cars. For cognitive distraction, the authors achieve 68% on the truck experiments and 86% for the passenger car experiments. However, the low number of drivers tested (3 for the passenger car, 12 for the truck) is insufficient to discuss the generalization capability of the system. In [29] several features related to the coordinates of 22 facial landmarks and driving data were used to predict accidents. In [30] the arm position, eye closure, eye gaze, facial expressions, and orientation provided by Kinect to detect visual and manual distraction on 6 subjects. All of these methods are different from our proposed approach either since they are not completely non-invasive or they have not been evaluated on such a large database.

The closest approach to the one that is presented in this article is the one by Li et al. [31], where the authors use AUs, gaze and head pose information to detect visual and cognitive distraction. Apart from the difference in the methodology to induce the cognitive distraction (that involves the drivers speaking), a very important difference in their approach is that they ask human evaluators to extract sequences of distraction based on the behavior of the drivers. This fact makes their work and ours incomparable, as the problem they try to tackle becomes how to detect human perception of expressions of distraction. However, it provides us a list of AUs related to this problem, which can be used for comparison. With the experiments performed on 20 subjects, the F-score for cognitive distraction detection is 79.4%.

To the best of our knowledge, ours is the first work that presents a completely automatic system that can be integrated in a vehicle, to detect the presence of cognitive load with an objective ground-truth, using non-intrusive visual monitoring of the driver's face and tested on such a large variety of subjects.

3 The EPV-DIST Database

One of the main contributions of this work is the introduction of a new video database with induced distraction during driving. In this section we describe the details of the database, which we name EPV-DIST, which is short for the EPFL-PSA-Valeo NIR Multi-Camera Database of Visual and Cognitive Distraction during Driving. The aim of the database is to provide videos of natural behavior of drivers while performing additional visual and cognitive tasks. This article focuses on facial expressions of cognitive distraction, therefore the visual distraction part is only briefly discussed. Another point worth mentioning is that the recording setup is built in a way that can be integrated directly in an actual consumer vehicle (in terms of camera positions), and provides robustness against real-life driving conditions, such as ambient light and head pose variations.

We have recorded 48 subjects, two of whom had to be excluded from the database due to technical problems during the recording. The subjects were recruited from students and research and administrative staff of EPFL and EPFL Innovation Park. As the mental tasks were prepared in French, they were asked to possess a sufficient level of understanding and speaking in French and also have a sight enough to drive without glasses, in order to avoid reflections of the NIR lighting. The subjects' ages are between 19 and 52 with an average of 30. The number of female and male subjects are equal and all subjects have given their consent for the use of their data in research on automatic visual behavior analysis. The length of the recordings is approximately 25 minutes per subject, making a total of more than 19 hours of recording.

The database will soon be publicly available solely for research purposes in the future, to help advance the research on facial behavior analysis during driving under various, predefined conditions. In the rest of this section we give details on the data acquisition setup and the experimental protocol, including the induction of the visual and cognitive distraction conditions.

3.1 Data Acquisition System

The EPV-DIST dataset consists of multi-view videos that are recorded using three NIR cameras and a special lighting equipment per camera with adequate filters, in order to filter out ambient light. Figure 1 shows the recording setup and the position of the recorded subject during the experiments.

(a) Left Camera (b) Frontal Camera (c) Right Camera (d) Generated Virtual Frontal View

Fig. 2. Images acquired by the three cameras and the corresponding generated virtual frontal view.

Both the choice of the recording material and the placement of the cameras (see Fig. 1a) are based on a realistic application in a real consumer vehicle. Since the light conditions change very often during driving we have performed the recordings with cameras with a wide wavelength capture range (PointGrey Flea3) and adequate band-pass filtering. We have also built three integrated NIR-LED (850 nm) circular lighting circuits that can be placed around the cameras and illuminated in synchronization with the frame-grab of the cameras using a microcontroller. These lighting systems make sure there is constant illumination around the face and the bandwidth filter, at the same wavelength as the LEDs, filters out a substantial amount of the ambient light. This allows for a continuous visibility of the driver's face with close to constant illumination and is suitable for a real application in a car in any light condition, e.g. when there is too much sunlight or while passing through a tunnel. The camera-light pair has already been used in our previous experiment on posed expressions of stress [32], where we have shown the feasibility of facial expression recognition with such a system in the single camera case. In addition to the circular LEDs, we have placed similar lighting at the four corners of the simulator screen which are illuminated in turns with the lighting around the cameras, to create the glint and dark/bright pupil effect for use in future research for gaze analysis. The gaze related features have not been included in this paper in order to restrict the focus on facial actions and to avoid possible bias and noise in the experimental results due to unoptimized gaze measurements.

As for the choice of the number and positions of the cameras, there exist two constraints for a realistic setup. The first one is the amount of head-pose coverage using a virtual frontal view generation from all cameras. The second one is the feasibility of placement of the cameras in an actual car, without blocking the driver's sight and where there is already support to place the camera. We have chosen to use a three-camera system and placed the cameras as can be seen in Fig. 1a. The first camera, the semi-frontal one (referred to as the frontal camera through the rest of the paper) is placed in the representative position inside the console behind the wheel. The left camera (with respect to the driver) is where would be the highest point of the left pillar in a car. Finally, the right camera is on the representative position of the bottom-left corner of the rear-view mirror.

The three cameras record frames synchronously at a rate of 20 fps as seen in Fig. 2a, 2b and 2c for the left, right and frontal cameras, respectively. We only use one out of two consecutively recorded frames (i.e. 10 fps), since the other one corresponds to the dark pupil frame, as explained previously. The three images from the three cameras are then used to reconstruct a virtual frontal view of the driver's face as seen in Fig. 2d. The details of this reconstruction are given in Subsection 4.1. This three camera system allows invariability against head pose changes, which occur quite frequently while driving, and also against occlusions that occur in one or more views. All these properties of the setup provide a realistic sense into our database, as it would be a suitable setup to integrate in an actual vehicle.

3.2 Experiment Protocol

The driving task we used for our experiments is the Lane Change Test (LCT) [33]. LCT is a simple, easy to manipulate driving simulator that has become the standard simulator for testing secondary tasks while driving [34]. It has been commonly used in the past for experiments involving such secondary tasks ([35], [36]). We have used a Logitech G27 wheel and pedals set for the control. The LCT allows continuous recording of the wheel and pedal motions, which are useful in providing a metric for the driving performance (explained later in detail).

The LCT consists of a series of lane change tasks which are presented as road signs on the simulator screen (see Fig. 1a) and the drivers are asked to change their lane according to the sign presented, as soon as they see the sign and as quickly as possible before passing by the sign. We have fixed the maximum speed at 60 km/h and the distance between two signs at 150 meters, which results in lane change sequences (LCS) of approximately 9 seconds, since the drivers were asked to maintain the maximum speed. The road signs appears $\sim$ 1 sec after the introduction of the distraction (if it applies), and disappear 40 meters later, giving the drivers around 2.4 seconds to perform the appropriate lane change. We use the whole LCS in our analysis as it simulates conditions that occur frequently during driving, such as changing the lane or keeping the right one with or without induced distraction.

3.2.1 Driving Conditions

All subjects were asked to perform the driving task in three conditions: The baseline, solely the driving task without any extractors; visual distraction, a visual secondary task

which requires the driver to take the eyes off the road; and cognitive distraction, where the attention is directed to a non-driving related task without the need to take the eyes off the road. All factors other than the distractive agents were kept the same for the three conditions. Each subject has completed a total of five driving tasks, three baselines, one visual distraction and cognitive distraction. The order of the distraction related tasks have been randomized among subjects, such that there is an equal number of subjects who have performed the visual task before the cognitive one and vice versa. This randomization is for decreasing the secondary effects of uncontrollable confounding factors, such as fatigue or disengagement.

The baseline condition is to obtain a ground measure for the driving performance and facial behavior without workload of the subject. It is performed three times in total, the beginning (after the familiarization part, which is not recorded), between two distractive conditions and the end. Each one consists of 18 lane changes, equally distributed for the 6 possible types of lane change between the left, right and middle lanes, in order to avoid effects of learning. In the end, we obtain 54 LCS, 9 for each lane change type, per subject.

The visual task used to induce visual distraction is the Surrogate Reference Task (SURT), which required subjects to look at a secondary screen on their right and therefore divert their visual attention from the road (Fig. 1a). The visual distraction part of the experiment has not been included in this article to keep the focus on cognitive distraction. Therefore, we only briefly introduce it and leave the analysis of the recorded data as future work. Yet, this part has been included in the database to provide the research community videos of varying head poses during driving, with a measure of the driving performance.

The final driving condition is the induced cognitive distraction, which forms the main focus of this article. This was performed using an auditory version of the Operation SPAN (OSPAN), developed by Turner and Engle [37] and has been used by the National Highway Transports Safety Administration in US (NHTSA) as a standard task simulating driver cognitive distraction. The OSPAN task makes use of the working memory and attention of the driver and does not require the visual attention as the task is presented in audio and the response is either manual or by speech.

The OSPAN task is composed of two components, the first one is making simple calculations and the other memorizing words. At each LCS the driver is told a simple mathematical calculation statement, e.g. *Two times four plus one is ten*. The driver needs then to press the corresponding hand pedal behind the wheel if they think the statement is true or false. The choice of pedal for the right and wrong answers have been randomized to reduce the effects of the natural tendency to unintentionally think that one side represents the correct one. Right after the statement the subjects also hear a simple word in French, e.g. *maison (house)*, *rouge (red)* or *chemise (shirt)*, which they were asked to memorize and repeat at the end. The LCS that we analyze do not include the part where the drivers repeat the words they had to memorize. In the easy condition the participants hear two mathematical calculations only including addition and subtraction along with two words to memorize, while in the hard condition they hear three calculations that also include multiplication and three accompanying words. All participants receive an equal number of easy and hard tasks following each other and the order of which has been randomized equally among participants. The calculations and words have been recorded prior to the experiment and was repeated from a speaker in the experiment room, synchronized to appear at the same instant for every subject. The OSPAN task creates an additional load to the working memory of the subject and aims to pull the attention of the driver off the road and the driving task. Note that we have not aimed at any positive or negative valence effect of the cognitive distraction, for example as performed in [38] by selecting words related to positive and negative emotions. It is also worth mentioning once again that, since we do not have any ground truth for whether the driver is actually distracted or not, our automatic system aims at predicting the cognitive load, which we know that exists during the corresponding driving condition.

In order to put the driver in a multitasking condition each distraction task started a couple of seconds before the appearance of the lane change sign. The participants had not been given any instruction prior to or during the experiment regarding the priority of the driving vs. secondary task. Each participant, therefore, chooses such a priority depending on his/her own workload and sometimes in a varying manner for each task, as observed from their recorded data.

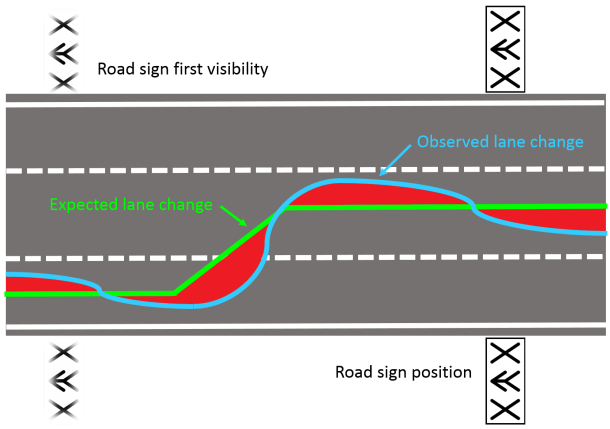
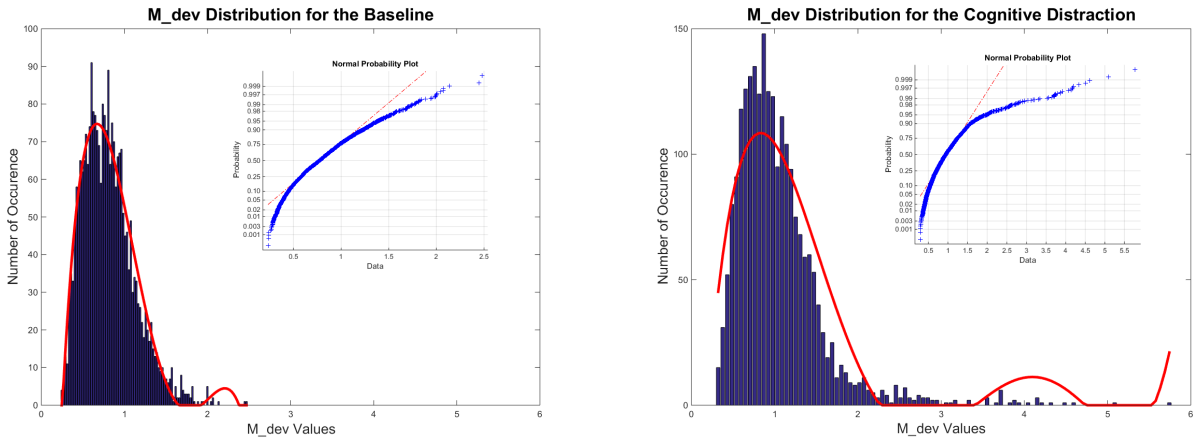


Fig. 3. Mdev calculation as the area between the expected lane change and the observed driver behavior

3.2.2 Measuring Driving Performance

The LCT simulation system allows recording the wheel and pedal motion at all times, which we then use in order to calculate a performance measure for each of the LCS. The measure that we calculate is the Mean Deviation from the normative lane change behavior, or Mdev. The Mdev is measured by calculating the area between the expected driving behavior for a specific lane change and the actual one (Fig 3). It is a standard way of quantifying the driving performance on a simulator over short distances [34]. The area between the two trajectories is sensitive to perception (missing the sign), reaction time, quality of the maneuver and lane keeping [33]. Note that the Mdev could also have been used as an indirect measure of the level of cognitive



(a) Distribution of the baseline M_dev values

(b) Distribution of the cognitive M_dev values

Fig. 4. Comparison of the M_dev values for the two driving conditions, with a polynomial fit plotted on top of each distribution. The smaller plots show the deviation from a normal distribution, indicating that the distributions cannot be assumed normal.

distraction, and a ground-truth for the classification problem in this article. However, due to the distribution of the Mdev values, this would have caused an imbalanced classification or regression task. Also, the problem we address in this work is to detect when a driver is imposed an additional mental load. Thus, we are interested in the facial behavior during situations that might cause *unsuccessful* driving, and not necessarily during unsuccessful driving. Therefore, we only use the Mdev values to show whether the induced cognitive distraction had an overall effect on the driving performance.

In order to show the effectiveness of the cognitive distraction induction that we used, we have performed a statistical analysis of the Mdev performance values, comparing the baseline and the cognitive task. We have calculated the Mdev for each 8.5 seconds sequence (the first 0.5 seconds were removed to remove noise) and Fig. 4 shows the distribution of the Mdev values among all sequences for every subject in the two conditions. The initial observations are the difference between the baseline (BL) and cognitive distraction (COG) in variance (0.098 for BL vs. 0.295 for COG) and the shift in the mean value (0.809 for BL vs. 1.048 for COG), median (0.772 for BL vs. 0.944 for COG) and the maximum values (2.481 for BL vs. 5.769 for COG). In addition, we have performed a Wilcoxon signed rank test on the average Mdev values of the subjects in the two conditions. The Wilcoxon test is a non-parametric paired difference test used to compare two related samples [39]. It is used to compare ordinal random variables that are non-Gaussian distributed, which fits perfectly our case (Fig. 4a and 4b). The signed-rank test gave a $p - value < 0.01$, showing that the two distributions are significantly different from each other, proving the effectiveness of the manipulation for the cognitive distraction. We also note that we observe a higher mean Mdev value for all 48 subjects, thus showing a worse performance under cognitive distraction.

4 Detection of Cognitive Distraction Segments - Methodology

This section describes the methods built and adopted in order to detect the presence of cognitive load via the driver's facial actions. In our context, this means classifying each recorded LCS as belonging to the baseline or cognitive task, as explained in Section 3.2. The outline of the pipeline is as following: First we generate a virtual frontal view of the driver's face in each frame using a Bilinear 3D face model and texture mapping from a 2D image. Then we detect 14 AUs on the generated virtual frontal view of the face by extracting Scale Invariant Feature Transform (SIFT) features and applying SVM classification for each AU separately. Next, we extract features from the dynamic continuous valued output of the SVMs, also investigating the correlated behavior between the AUs and finally feed these features in an SVM or Random Forest classifier to obtain a binary response for each sequence as distracted or not. The details of each method, as well as their implementation are given in the rest of the section.

4.1 Virtual View Generation from Three Cameras

The model based face pose normalization / frontalization has been applied widely in face recognition [40], [41], [42]. It is also known as virtual face frontal view generation. One can fit a 2D deformable mesh model to a non-frontal face and apply non-linear warping to generate a virtual frontal face [40]. However, it has been shown that warping with a sparse 2D mesh model is sub-optimal due to artifacts and discontinuity. Instead, we fit a 3D dense mesh model and map the texture directly to the mesh vertices. The frontal view face is rendered by applying inversed rigid motion of the 3D face model. Fig. 5 shows the concept of our face pose frontalization method on an example non-frontal face image and the resulting transformation.

Fitting a 3D dense face mesh model with texture information is far from efficient for real time application. We adopt a feature based 3D mesh model fitting whose fitting efficiency and accuracy are good enough for real

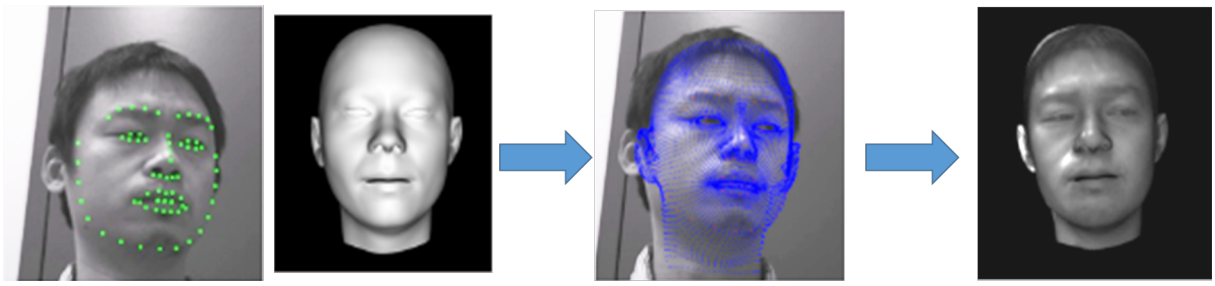


Fig. 5. Virtual face generation pipeline and the resulting image.

time virtual frontal view generation. In order to recover the expression variations and identity variations of human faces for emotion detection applications, we employ a bilinear 3D morphable model [43]. The model has two sets of parameters to control expression changes and identity changes separately.

The objective of the 3D model fitting is to minimize the projection error of the facial landmark features, with respect to a set of corresponding 2D features detected on a 2D facial image. In total 68 salient facial features are selected and the 2D salient facial features are detected and tracked using Supervised Descent Method (SDM) [44], which is reused in the AU detection step as explained in Sec. 4.2.

The feature based 3D face mesh model fitting can be easily extended to multiple camera setup. The coefficients of the bilinear face model are not dependent on the viewpoint because they are characterizing a 3D object's shape and not its projection on the image plane. It has been shown in [45] and [46] that fitting a 3D morphable model in a multi-view setup provides more accurate and robust results. Therefore, we reconstruct the mesh based on the tracking from the three cameras positioned as shown in Fig. 1a.

To generate a virtual frontal view image, we extract the texture information from a 2D image and map the values on the corresponding vertices of the reconstructed 3D face mesh. The texture information can be extracted from a specified camera view, or the optimal camera view, or an adaptive fusion of multiple camera views. In this work, we obtain the pixel values from the view with the smallest absolute yaw angle, which is mostly the frontal view due to the nature of the driving and secondary tasks. Given a reconstructed 3D face mesh $\mathbf{f}$ and its estimated projection operator $\mathbf{L}$ with respect to an input face image $\mathbf{I}$, the visible vertices in $\mathbf{f}$ are determined by using the normals and the viewing angle. Those vertices are projected on the 2D image plane with the projection operator and the underlying pixel values T are assigned to their corresponding visible vertices. An example of a rendered frontal face image is show in Fig. 5. Fig. 2 shows the three views and the reconstructed virtual face on an highly expressive real-case frame from our database.

4.2 AU detection from Virtual Frontal View

Once we generate the virtual frontal view from the three cameras, we detect 14 AUs from the generated frame. For this purpose we adopt the system that we have recently proposed [15] and that has won the AU occurrence detection sub-challenge of the FERA2015 [47]. The FERA2015 challenge was organized to promote advances in research on AU and AU intensity detection. It is composed of two challenging datasets (BP4D [48] and SEMAINE [49]) with spontaneous and natural behavior each annotated frame-wise for the presence and intensities of AUs. The participants were provided with two sets of training and development partitions and asked to send a working program that would be applied on two unseen test partitions, in order to assess the efficiency of the systems in a blind manner, i.e. without the advantage of parameter tuning or usage of prior knowledge on the data. Our framework presented in the scope of FERA2015 allows us to obtain a continuous AU occurrence signal for 14 AUs, which are listed in Table 1 along with their definitions. An overview of the system is presented here and for more details the reader is referred to the related proceedings article [15].

TABLE 1
Detected AUs and their definitions

Action Unit	Definition
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU6	Cheek Raiser
AU7	Lid Tightener
AU10	Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU23	Lip Tightener
AU25	Lips Part
AU28	Lip Pucker
AU45	Blink

The initial step in the AU detection system is to locate the facial landmarks, around which we will then acquire the relevant appearance based features. For this purpose, we use the state-of-the-art face tracker based on SDM [44]. The SDM starts with an initial guess and estimates the shape using a cascade of regression models that are learned at each step using local texture features (e.g. SIFT) extracted from the landmarks estimated in the previous step. Note that, since the virtual view generation and AU detection systems are currently implemented as two separate pipelines, we reapply the SDM tracker on the generated virtual view. In

the future, these systems will be combined for efficiency reasons. The SDM outputs the locations of 49 landmarks and using this mask we calculate the locations of 8 additional *non-salient* landmarks. The details for the calculation are present in [15]. These additional points (AP) are generally excluded from face trackers or facial landmark detectors as they mark transient features of the face and their annotation and detection are not as trivial as the non-transient landmarks. However, they contain very important local appearance information related to facial actions as many appearance changes occur around these points during certain muscle contractions. These points can be seen on an example virtual face image from the EPV-DIST database in Fig. 6 along with original SDM landmarks. Their locations and some of the AUs they are related to are listed as follows:

- AP1 - The center of the eyebrows, relevant to AU4 and AU1
- AP2 and AP3 - Around the crow-feet wrinkles, relevant to AU6 and AU7
- AP4 and AP5 - Sides of the nose, relevant to AU10 and AU9 (nose wrinkler)
- AP6 and AP7 - Nasolabial furrows, relevant to AU6 and AU10
- AP8 - On the chin, relevant to AU17

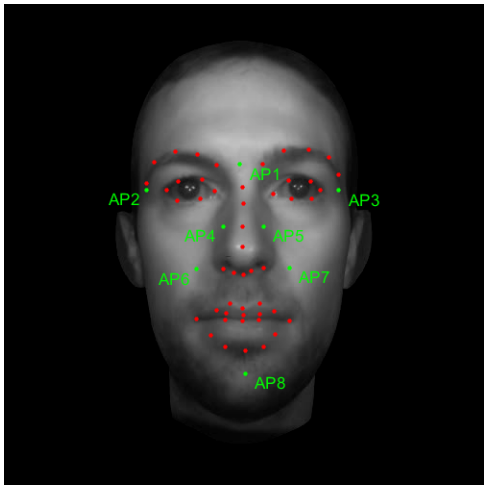


Fig. 6. The facial mask used to obtain appearance features. The red points show the original SDM landmarks and the green ones are the additionally calculated points.

After obtaining the landmarks from the face tracker, the face is aligned using the eye locations to correct for any possible in-plane rotation still remaining from the virtual view generation. This is performed before the APs are extracted, so that the calculation of their locations is invariant to the head-pose. Correcting only for the in-plane rotation is sufficient since the training data we have used consists of mostly frontal faces and the system is applied on virtual frontal faces. Later, the face is scaled to a fixed size of 200 by 200 pixels and the SIFT features [50] are extracted around the 57 landmarks in total. SIFT features have been effectively used in mainly object recognition and tracking ([51], [52]) and successfully applied on the AU detection problem as well [53], [54]. The SIFT descriptors extracted in a 32 by 32 local neighborhood around each landmarks result in a feature vector of size 7296, which is then reduced using Principal Component Analysis (PCA), retaining a certain number of final features learned for each AU separately.

TABLE 2
F1-Scores on the FERA Challenge

Database	BP4D	
AU	Prop. System [15]	Best Baseline [47]
1	**0.261**	0.188
2	0.167	**0.185**
4	**0.283**	0.197
6	**0.729**	0.645
7	0.785	0.799
10	**0.802**	0.801
12	0.779	**0.801**
14	0.625	**0.72**
15	**0.348**	0.238
17	**0.380**	0.311
23	**0.441**	0.320
Average	**0.508**	0.473
Database	SEMAINE	
AU	Prop. System [15]	Best Baseline [47]
2	**0.655**	0.569
12	**0.769**	0.595
17	**0.215**	0.091
25	**0.623**	0.445
28	**0.251**	0.250
45	0.325	**0.396**
Average	**0.481**	0.391

These features are used to train a L1-regularized linear-SVM classifier for each AU separately on a custom made training set that includes images from the CK+ [55], GEMEP-FERA [56] databases in addition to the challenge datasets SEMAINE [49] and BP4D [48]. The training set consists of 6713 images in total and, in addition to the well-established standard database CK+ of posed expressions, includes many non-posed, or spontaneous, examples of expressions from the other three databases. This fact is particularly useful when the system is applied on real data, as in the case of our application.

The results we obtained on the unseen test-set of the two challenge datasets are shown in Table 2 in comparison with the best challenge baseline results. The presented F1 scores on this challenging AU detection problem shows the efficiency of the system and proves suitable for use in a real application. Note that, although the original article [15] proposes a multi-label manifold embedding scheme to improve AU detection and achieves a better result on one of the two unseen partitions, we have chosen not to adopt this part of the system in order to obtain a better generalization on unseen data. Table 2 shows the results obtained using the system applied for this article, that is without the multi-label manifold embedding.

The SVM classifiers each give a continuous value output, which is the distance to the hyper-plane. It has been long debated in the community whether the output of classifiers trained in a binary manner should be used to quantify the intensity of AUs. For example, recently Girard et al. have shown that the intensity of smiles are better recognized using classifiers directly trained with annotated intensities [57]. Nonetheless, we use the decision to the hyper-plane of the SVM as a relative intensity measure since it provides enough comparative information when the purpose is not a

direct AU intensity detection defined by the FACS [14].

4.3 Feature Construction

For the classification between the sequences belonging to the baseline and cognitive distraction we extract features from the AU signals obtained using the system described in 4.2. The signals contain the SVM distances to hyperplane at each time instance, which are representing the AU intensities. The sequences that contain too few frames due to errors during recording or face detection failure because of heavy occlusion (e.g. by the hands on top of the steering wheel) have been removed from the analysis resulting in a total of 4520 LCS. In total we have removed 172 segments from the analysis, but the removed segments were distributed almost equally among subjects and driving conditions, since most of them corresponded to the last sequence of each driving condition. These sequences had to be removed due to a synchronization problem in order not to bias the results and have the same length of LCS for all data points.

The first set of features, which we will refer to as Feature Set 1 from this point on, come directly from the continuous AU signals. For each of the 14 AU signals (see Table 1 for the list) we obtain the mean, variance, maximum and minimum values along the 8.5 second sequences. This process is performed by dividing the sequence in four in time. The reason for splitting the signals in time is to make use of the differences in AU behavior that may occur on different portions (or quarters) of the LCS. For instance, a person might display a facial reaction while listening to the calculation sentence he/she needs to respond to, or during the lane change task which follows the auditory input. Splitting the feature extraction into smaller segments makes it feasible to extract this sort of dynamic information and splitting them in four is suitable in terms of the ordering of the driving task and cognitive distraction induction.

The second type of features (Feature Set 2) are derived from the cross-correlations of AUs on different time delay levels. While constructing these features we were inspired by the Appraisal Model of Emotion, as proposed by Scherer [58], which states that the activation of certain physiological components are coupled, or synchronized, when we are faced with an emotional stimulus. Also following this theory, Kroupi et al. have shown coupling between the phase and amplitude of the EEG and EDA signals while the subjects are watching emotionally stimulating music videos [59]. Another example of a similar analysis is the multiple works by Williamson et al., who have shown the existence of a difference in coordination, movement, and timing of vocal and facial components between patients suffering from major depressive disorder (MDD) vs. control subjects [60], [61], winning the AVEC 2013 [62] and AVEC 2014 [63] challenges on automatic detection of MDD severity.

Using a similar idea, we calculate the cross-correlation between each of the 14 AUs, within a delay of -80 to $+80$ frames with 2 frames interval. This corresponds to a signal of length 81 for each AU pair and allows modeling the sequential behavior between AUs on a scale of -4 to $+4$ seconds. From those signals we extract, once again, the mean, variation, maximum and minimum values, in addition to the location in time of these maximum and minimum values, and the correlation values at delays corresponding to $-40, -30, -20, -10, 0, 10, 20, 30$ and 40 frames, i.e. at each second in a bi-directional manner. This enables us to obtain an extensive set of features that represent factors like the total amount of co-activation and its variation, moments of maximum and minimum synchronization and the level of co-activation at certain levels of delay between AUs located in similar or different parts of the face. Finally, we truncate the feature set according to the correlation priors between AUs. This truncation serves for keeping AU combinations that frequently occur and removing those with little or no correlation. In case such an unusually high correlation is observed, for instance caused by a distortion from the virtual view reconstruction due to heavy head-pose, this process will make sure this noisy observation has no effect on the overall feature set. As correlation priors we use the co-occurrence table of AUs obtained from the AU detection training set, as used in [15], and use a threshold of 0.25 as shown in Fig. 7. This ratio was chosen empirically and allows for discarding the AU pairs that are not *naturally* and commonly related to each other.

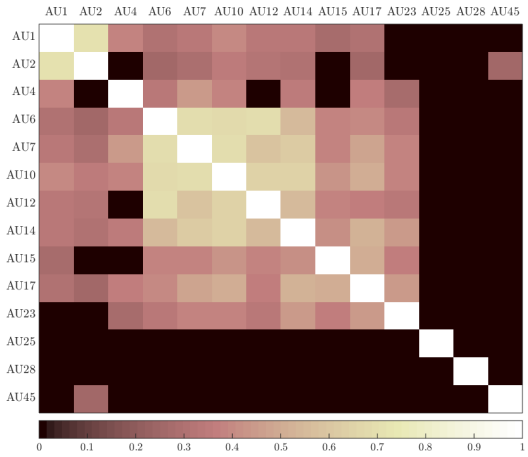


Fig. 7. Correlation table for the 14 AUs, a higher correlation indicates a high number of co-occurrences between AUs in the training set. The matrix was truncated for values < 0.25 to indicate the cross-correlations that were excluded

Our hypothesis is that this dynamic co-activation information will help better differentiate the facial behavior of the complex mental state that is cognitive distraction. In Section 5, we show that, indeed the cross-correlation based features improve the accuracy on a subject based analysis, yet they are not so helpful for the subject independent classification task.

4.4 Person Specific Normalization for Classification using SVM and Random Forests

The final component of the distraction detection system is the classification part. For this, we use linear SVM for the subject based tests, where the training and test examples are relatively on similar manifolds compared to between subject tests. For the subject independent tests, we therefore compare the performance of the SVM with Random Forests (RF) classifiers. RF are known to be less effected by overfitting thanks to their bagging mechanism [64]. They learn

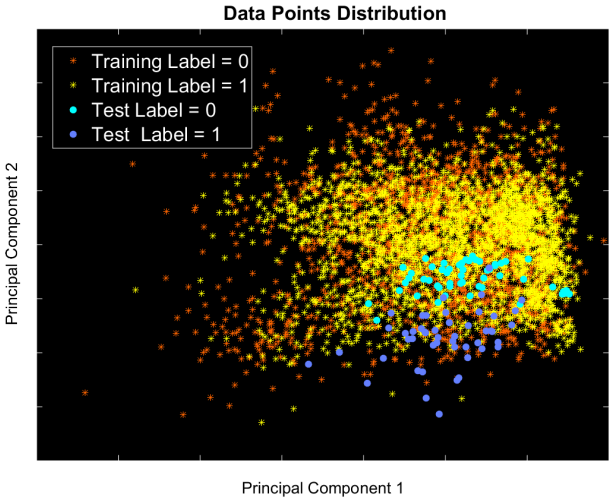


Fig. 8. The data distribution on the two first principal axes, with an example subject chosen as the test case. Label = 0 for baseline and 1 for cognitive distraction condition. (better visualized when printed in color)

the best splitting by multiple features each time randomly choosing a random subset of samples and features. They are also more suitable for cases with a large number of features.

For the SVM training and testing we make use of the LibSVM library [65] and for the RF we use the Scikit-learn machine learning library for python [66]. The hyper-parameter C for the SVM and the number of trees, maximum number of features and minimum number of samples per split hyper-parameters of the RF were optimized using a 5-fold cross-validation on the training data in a subject independent manner.

While analyzing the data we have discovered that although both type of features are very effective in discriminating the distraction and baseline sequences of individual subjects (see Subsection 5.1) the performance on the subject independent tests are very low. We assume that the reason is that the types of features that we use are discriminative enough to model individual behavior, yet they remain too person-specific. Indeed, visualizing the data, we have seen that most of the subjects are clustered among their own samples. Fig. 8 shows an illustration of this phenomenon using the first two principal components of Feature Set 1 after PCA applied on all training data. We can see that even using two dimensions the data points belonging to the test subjects can be easily separated into the labels, yet the same is not true for the training data, for which samples from the two labels do not demonstrate any noticeable pattern and are scattered across the feature space instead. Of course, the projection on 2D is not very meaningful when using complex classification methods; Fig. 8 is only for illustrative purposes.

To overcome this problem, instead of the common convention of normalizing the whole training data to zero-mean and unit standard deviation, we propose to perform this operation subject wise, as:

$$\forall x \in F : x = \frac{x - \overline{x}_s}{\sigma_s} \quad (1)$$

where x is any point in the features set F belonging to the subject s, $\overline{x}_s$ is the mean of all data points belonging to subject s and σ_s the standard deviation. Even though, this may seem as a factor preventing a real-time application on an unseen subject, the only implication it brings is actually the need for some seconds of frames from the considered subject. In other terms, the person based normalization is completely unsupervised, does not require any re-training of the classifier (as we only need to change the placement of the test subject) and as seen in Section 5.2 increases substantially the subject independent detection rates with ~ 100 data points per subject.

5 Results

This section presents our classification result for the baseline vs. cognitive distraction cases. Out of the 4520 LCS in total (~ 100 per subject), the number of the sequences for the cognitive distraction case is 2156. We present our experimental results for the classifiers trained per-subject and in a subject independent manner using different feature configurations and classifiers. Table 3 presents the accuracies for the best performing systems for the two types of experiments, serving as a summary of the results and the details are presented in the rest of the section.

TABLE 3
Results (in percentage) of Best Performing Systems for Subject Independent and Dependent Cases - OA: Overall Accuracy, F1: F-score, Prec.: Precision, Rec.: Recall

	OA	F1	Prec.	Rec.
Sub. Dependent	95.51	95.16	96.38	93.97
Sub. Independent	68.10	65.57	67.22	64.00

5.1 Subject Dependent Cognitive Load Detection

We first train classifiers independently for each of the 46 subjects in a leave-one-out manner. That is, we learn the classifier hyper-parameters using a 5-fold cross-validation and train the classifier with the best parameters on all sequences for a certain subject points except for one, and test it on the left-out sequence, or data point. The classifier we use for the subject dependent tests is the linear SVM. In Table 4 we show the results obtained using Feature Set 1 alone, Feature Set 2 alone and the two in combination, and compare the accuracies obtained with and without the truncation of correlation features as explained in Sec. 4.3. We have also performed a Wilcoxon signed rank test on the per-subject accuracy distributions of each pair of methods. The tests resulted in a $p - value < 0.01$ each time, showing the statistical significance of the comparison of the type of features.

As shown in Table 4, the best results are obtained by combining the features extracted directly from AU signals (Feature Set 1) and those from the cross-correlations (Feature Set 2), supporting our hypothesis that the dynamic inter-relations of AUs are useful in determining individuals' expressions of cognitive load. Using Feature Set 2 alone also proves as efficient as using Feature Set 1. The best accuracies obtained are 95.51% with a standard deviation (std.) across

TABLE 4
Subject Dependent Detection Results - OA: Overall Accuracy, F1: F-score, FT: Feature Truncation on Set 2

Feature type	OA (%)	F1 (%)
Feature Set 1	93.74	93.39
Feature Set 2	93.85	93.47
Feature Set 2 + FT	93.89	93.49
Features Sets 1 + 2	94.88	94.57
Feature Sets 1 + 2 + FT	**95.51**	**95.16**

subjects of 3.44 for the overall accuracy and 95.16% for the F-score with std. 3.67. These values are calculated over all data points, which corresponds to an average weighted by the number of sequences per subject. The very high accuracy measures, and low variation among subjects, demonstrate the efficiency of the proposed system, when it is trained on labeled data of a specific subject. A side-observation is that, the feature truncation improves accuracy in both of the relevant cases (Feature Set 2 alone and Feature Set 1 and 2 combined), validating the usefulness of exploiting prior AU co-occurrence information. The per-subject accuracies for the best performing system are shown in Fig. 9, which will be referred to again in the following subsections.

5.2 Subject Independent Cognitive Load Detection

The second set of experiments we have performed is the subject independent tests, that is carried out in a leave-one-subject-out manner. This time, we also use RF in comparison with SVM, since RF are known to be less affected by over-fitting on training data, or subjects in our case. We have not included the results using RF in the driver dependent experiments as SVM gave better results by making the best use of the relatively low variance subject based datapoints. Table 5 presents the results obtained using both classifiers, Feature Set 1 and 2 alone and in combination, additional PCA (retaining 98% of the total variance, performed for SVM only since RF internally handle the problem of irrelevant features) and the subject based normalization as explained in Sec. 4.4. Similarly to Sec. 5.1, we have also performed a Wilcoxon signed rank test to statistically compare each pair of experiments and obtained a $p - value < 0.05$ for each one.

The best results are obtained using RF classifier with Feature Set 1 alone when the person-specific normalization is applied with overall accuracy 68.10% (std. = 12.71) and F-score 65.79% (std = 14.02). The person-specific normalization is indeed very effective with all features types, especially with RF. This confirms our rationale explained in Sec. 4.4, claiming that the data points of each subject are clustered separately in the feature space. However, it is not effective enough to obtain an accuracy close to the classifiers trained in a subject based manner (Sec. 5.1). As it can be seen in Fig. 9 this effect is more critical in certain subjects (e.g. Subjects 4, 7, 13) and less in others (e.g. Subjects 5, 6). Also, we observe that the correlation related features (Feature Set 2) do not increase the detection efficiency when used in combination with Feature Set 1, and also result in lower results when used alone. These results suggests the individuality of such

TABLE 5
Subject Independent Detection Results - OA: Overall Accuracy, F1: F-score, FS: Feature Set, SN: Subject wise data normalization

Cl. type	Feature type	OA (%)	F1 (%)
SVM	FS 1	63.36	59.69
	FS 1 + PCA	63.74	58.42
	FS 1 + SN	**65.5**	**62.32**
	FS 1 + PCA + SN	65.35	61.77
	FS 2	57.85	54.95
	FS 2 + PCA	59.38	55.31
	FS 2 + SN	61.39	58.95
	FS 2 + PCA + SN	62.72	62.14
	FS 1 + 2	61.82	59.19
	FS 1 + 2 + PCA	59.58	54.65
	FS 1 + 2 + SN	62.99	60.92
	FS 1 + 2 + PCA + SN	63.96	61.49
RF	FS 1	63.98	61.93
	FS 1 + SN	**68.10**	**65.79**
	FS 2	57.19	59.31
	FS 2 + SN	63.98	61.49
	FS 1 + 2	58.83	60.61
	FS 1 + 2 + SN	65.29	64.17

dynamic multi-AU patterns, i.e. that this kind of information is more meaningful when it is learned on each subject independently. This problem of individuality is discussed further in the rest of the paper.

5.3 A look into the relevant features

In order to see which AUs or AU pairs are the most relevant to our proposed classification task we inspect the correlations of each feature in Feature Set 1 and 2 with the ground-truth labels for baseline and cognitive distraction segments. Since the subject dependent classification is significantly more efficient compared to the subject independent one, we find it more rational to perform this analysis on a subject level as well.

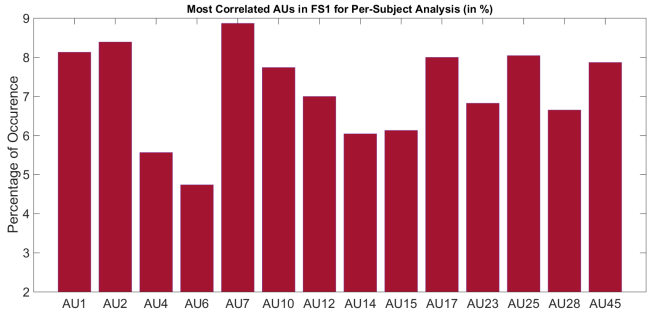


Fig. 10. Percentage of each AU within the 50 most correlated features for each subject in Feature Set 1

First, we calculate the correlation of all 224 features from Feature Set 1 with the labels for each 46 subjects. Then, for the 50 most correlated features for each subject we look at which AU signal and which temporal segment they belong to. Fig. 10 shows the total percentage of each AU among

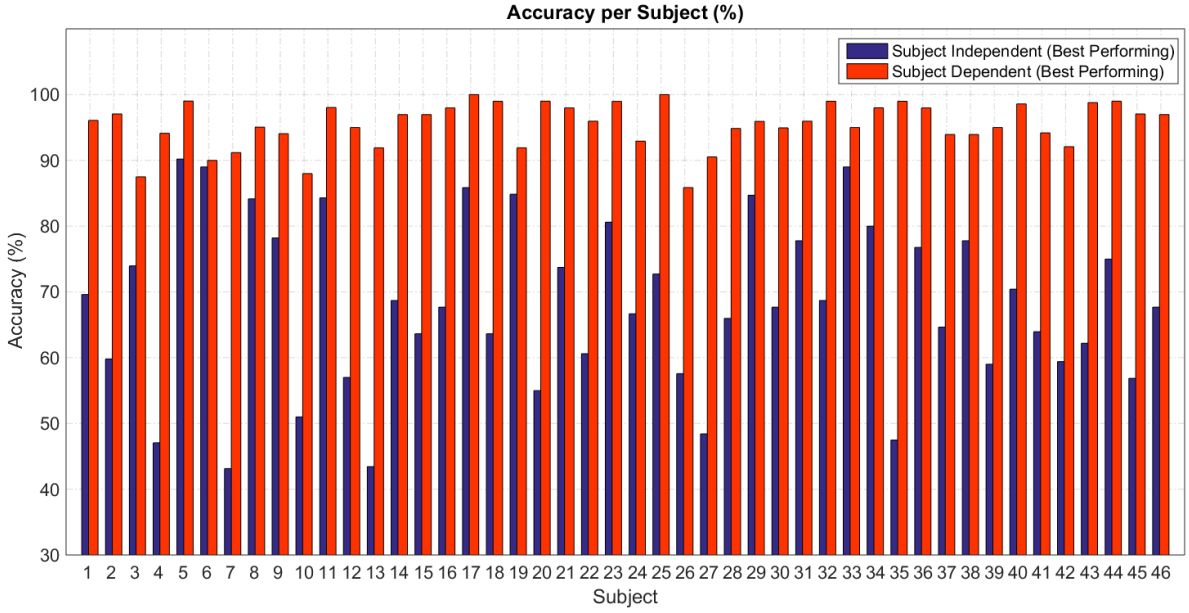


Fig. 9. Overall Classification Accuracies for each subject, for the best performing methods in subject independent and subject based training conditions

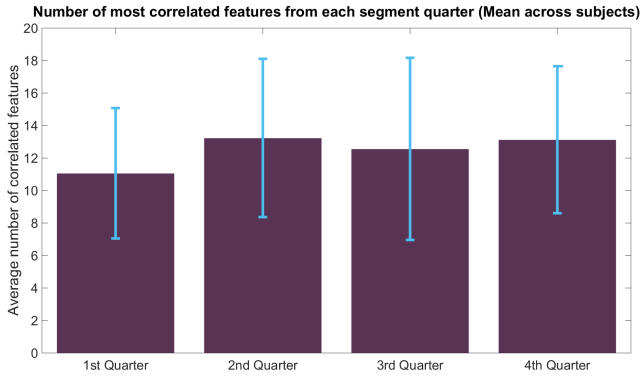


Fig. 11. Average (and std.) number of features selected from each time segment of the sequences within the 50 most correlated for each subject in Feature Set 1

those features, and Fig. 11 shows the mean and standard deviation for each of the four segments (std. removed from Fig. 10 for clarity of presentation). We observe that the AU that appears most in the analysis is AU7 (eye-lid tightener), which indeed appears frequently in expressions related to concentration, thinking or focusing. It is followed by the outer and inner eye-brow raise motions AU2 and AU1, lips part AU25 and chin raise AU17, without any clear difference in amount of occurrence. The fact that many AUs occur frequently in the list of correlated features once again shows the large variety of expressions related to cognitive load, and helps explaining the difficulty in obtaining a highly accurate subject independent system. Two of the five most correlated AUs (AU1 and AU17) are also in line with the features found relevant to human perception of cognitive distraction, reported in [31]. For the temporal segments, none of the segments seem to dominate the others; yet, the first quarter is observed to appear less. This is expected, since it corresponds to the first two seconds of the LCS where the secondary task is presented (mental calculation) and the lane change task appears only in the second quarter, forcing the driver to divide his attention and workload between tasks.

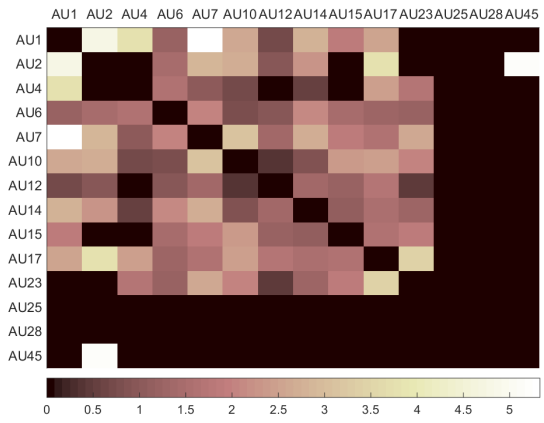


Fig. 12. Percentage of the most correlated AU pairs within the 100 most correlated for each subject in Feature Set 2

We perform the same procedure for Feature Set 2 and plot the percentage occurrence of features belonging to AU pairs as seen in Fig. 12. This time we investigate the 100 most correlated features, as the whole set is larger (of size 750). Some relevant AU pairs worth mentioning are $AU7 - AU1$, $AU45 - AU2$, $AU1 - AU2$, $AU1 - AU4$, $AU17 - AU23$ and $AU2 - AU17$. Although it is harder to interpret the features this time, we observe that most of them are related to eye / eye-brow actions, as in the single AU case. Interestingly, the most correlated upper AU pair includes the AUs identified

as related to fatigue in [22], which possibly implies an effort to regain attention, commonly in the two conditions. The lower face combination $AU17 - AU23$, on the other hand, can appear in expressions related to pensiveness or assessment of coping potential depending on the simultaneous upper face actions. Therefore, it makes sense that this combination is relatively meaningful for the differentiation between cognitive distraction and baseline.

5.4 Discussion

The presented results demonstrate that, although a subject based training or adaptation is necessary in order to obtain a highly precise detection, the subject independent system still achieves an acceptable accuracy in detecting the sequences with cognitive load. It is a known fact that the cognitive distraction is not one of the basic emotions (or states) that are conveyed similarly by everyone in terms of facial expressions. The subject dependency issue is therefore expected in systems aiming at recognizing such complex expressions. As for the unsupervised subject based normalization proposed, with an unseen driver the system requires only some couples of seconds of images of the driver's face to increase the detection rate $\sim 4\%$.

As stated earlier, the system is designed so that it can be completely integrated into a passenger vehicle. All components work in real-time ($> 15fps$) and integrating the two separately implemented pipelines of virtual view generation and AU detection will also increase speed.

A possible real-world application could be to integrate the system within the human-machine interface of the vehicle, and to activate a visual or audio alert to warn the driver in case a critical level of distraction is detected, as a part of Advanced Driver Assistance Systems (ADAS). With the semi-automatic driven cars slowly entering our lives, such systems gain even more importance, for instance to assess the driver's state when the driver needs to retake the car's control or to decide when it is safe (or suitable) to switch to fully autonomous driving. The current system outputs a decision based on 8.5 seconds of recording due to the definition of the lane change task, but is fully adaptable to shorter or longer durations and to the fusion of multiple sequences. For instance, a moving window that collects distraction information in time can be utilized and the relevant alert system could be activated when the number of segments involving cognitive load reaches a certain threshold. According to the detected level of cognitive load the severity of the countermeasure can also be adjusted, ranging from a small alerting beep or a message on the console to automatically slowing down or even stopping the car when conditions are suitable.

There are still many issues to handle for the real-world use of the system. Firstly, even though we propose a system with a NIR light and infrared system, images recorded outdoors may still differ in terms of color, texture and quality. We will perform outdoor recordings in order to assess the robustness of the system. A general convention in driver feed-back systems is to alert the driver timely, only when really needed and in a way that does not annoy the driver. This requires a good balance between the precision and recall, i.e. false positives and false negatives. Further user studies need to be performed in real driving conditions to assess the robustness of the detection, e.g. considering the head movements in real conditions, but the presented results already show the applicability of the proposed system. The system is also suitable to be adjusted for the precision/recall ratio (e.g. by tuning the decision level of the classifiers). The length of the temporal window, the four seconds delay introduced in feature construction and the previously mentioned threshold should also be tuned to obtain the best compromise between the driver comfort and safety.

Distraction does not affect our lives only in the driving context. Knowing that abnormalities in maintaining attention are symptoms of disorders such as Attention deficit hyperactivity disorder (ADHD), Asperger's syndrome or other Autism spectrum disorders, and considering the high accuracy of our subject dependent system, another possible use of the proposed system could be a personalized monitoring system to provide feed-back during treatments, that involve interaction with a human or a machine, of individuals suffering these disorders. A review of works on such interactive technologies can be found in [67].

6 Conclusion

We have presented a database, called EPV-DIST, of 46 people recorded using three cameras while driving a simulator in baseline, visual distraction and cognitive distraction conditions. The recordings have been configured to represent a configuration that could be integrated and work robustly inside an actual car during real driving conditions. Then, we have demonstrated a complete pipeline to discriminate the cognitive distraction segments from the baseline based on AUs. The proposed system first reconstructs a virtual frontal face image using the input from the three cameras, applies AU detection on the virtual image, then uses features extracted from the dynamic AU signals and cross-correlations of AU pairs to classify segments in the two driving conditions.

Using different configurations and methods we obtain an accuracy of $\sim 95\%$ when the system is trained separately on each subject, and $\sim 68\%$ in the subject independent case. Based on these results and further analyses, we identify that facial expressions of cognitive load vary hugely among subjects and also report the AUs and AU pairs that show relevance most commonly among the subjects. The completely automatic non-intrusive detection system is ready to be accommodated in consumer vehicles for use within applications aiming to prevent, or decrease, human error in accidents. Our further research will include the gaze and head-pose related features and their benefits for detecting the various types of distraction along with AUs. Compared to existing related work, this study is the one performed with the highest number participants using solely automatic analysis of facial actions and we hope the introduction of the database will stimulate further research in the field.

Acknowledgments

The authors would like to thank Estelle Chin and Olivier Pajot from PSA Peugot-Citroën Research and Patrick Bonhoure, Stéphanie Dabic and Julien Moizard from Valeo for

their invaluable contribution in our collaboration during the data acquisition and analysis of the data.

REFERENCES

[1] M. Pettitt, G. E. Burnett, and A. Stevens, "Defining driver distraction," in *12th World Congress on Intelligent Transport Systems*, 2005.

[2] J. D. Lee, K. L. Young, and M. A. Regan, "Defining driver distraction," *Driver distraction: Theory, effects, and mitigation*, pp. 31–40, 2008.

[3] M. A. Regan, C. Hallett, and C. P. Gordon, "Driver distraction and driver inattention: Definition, relationship and taxonomy," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1771–1781, 2011.

[4] J. C. Stutts, A. A. Association *et al.*, *The role of driver distraction in traffic crashes*. AAA Foundation for Traffic Safety Washington, DC, 2001.

[5] D. L. Strayer, J. M. Watson, and F. A. Drews, "Cognitive distraction while multitasking in the automobile," *Psychology of Learning and Motivation-Advances in Research and Theory*, vol. 54, p. 29, 2011.

[6] C. Galéra, L. Orriols, K. MBailara, M. Laborey, B. Contrand, R. Ribéreau-Gayon, F. Masson, S. Bakiri, C. Gabaude, A. Fort *et al.*, "Mind wandering and driving: responsibility case-control study," *BMJ*, vol. 345, p. e8105, 2012.

[7] C. Lemercier, C. Pêcher, G. Berthié, B. Valéry, V. Vidal, P.-V. Paubel, M. Cour, A. Fort, C. Galéra, C. Gabaude *et al.*, "Inattention behind the wheel: How factual internal thoughts impact attentional control while driving," *Safety science*, vol. 62, pp. 279–285, 2014.

[8] S. Bakiri, C. Galéra, E. Lagarde, M. Laborey, B. Contrand, R. Ribéreau-Gayon, L.-R. Salmi, C. Gabaude, A. Fort, B. Maury *et al.*, "Distraction and driving: Results from a case–control responsibility study of traffic crash injured drivers interviewed at the emergency room," *Accident Analysis & Prevention*, vol. 59, pp. 588–592, 2013.

[9] G. Berthié, C. Lemercier, P.-V. Paubel, M. Cour, A. Fort, C. Galéra, E. Lagarde, C. Gabaude, and B. Maury, "The restless mind while driving: drivers thoughts behind the wheel," *Accident Analysis & Prevention*, vol. 76, pp. 159–165, 2015.

[10] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," *National Highway Traffic Safety Administration, Paper*, no. 05-0400, 2005.

[11] R. L. Olson, R. J. Hanowski, J. S. Hickman, and J. L. Bocanegra, "Driver distraction in commercial vehicle operations," Tech. Rep., 2009.

[12] M. T. W. Victor, J. D. Lee, and M. A. Regan, *Driver Distraction and Inattention: Advances in Research and Countermeasures*. Ashgate Publishing, Ltd., 2013, vol. 1.

[13] K. L. Young and P. M. Salmon, "Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods," *Safety science*, vol. 50, no. 2, pp. 165–174, 2012.

[14] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, 1978.

[15] A. Yuce, H. Gao, and J.-P. Thiran, "Discriminant multi-label manifold embedding for facial action unit detection," in *11th IEEE Int'l Conf. on Automatic Face and Gesture Recognitions (FG 2015), FERA 2015 Challenge*, 2015.

[16] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 12, no. 2, pp. 596–614, 2011.

[17] H.-B. Kang, "Various approaches for driver and driving behavior monitoring: a review," in *Computer Vision Workshops (ICCVW), 2013 IEEE Int. Conf. on*. IEEE, 2013, pp. 616–623.

[18] W. Rongben, G. Lie, T. Bingliang, and J. Lisheng, "Monitoring mouth movement for driver fatigue or distraction with one camera," in *Intelligent Transportation Systems, 2004. Proc. of The 7th Int. IEEE Conf. on*. IEEE, 2004, pp. 314–319.

[19] H. Gu and Q. Ji, "Facial event classification with task oriented dynamic bayesian network," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proc. of the 2004 IEEE Comput. Soc. Conf. on*, vol. 2. IEEE, 2004, pp. II–870.

[20] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 11, no. 2, pp. 300–311, 2010.

[21] L. M. Bergasa, J. Nuevo, M. Sotelo, R. Barea, M. E. Lopez *et al.*, "Real-time system for monitoring driver vigilance," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 7, no. 1, pp. 63–77, 2006.

[22] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *Human–Computer Interaction*. Springer, 2007, pp. 6–18.

[23] F. Tango, M. Botta, L. Minin, and R. Montanari, "Non-intrusive detection of driver distraction using machine learning algorithms." in *ECAI*, 2010, pp. 157–162.

[24] M. Wöllmer, C. Blaschke, T. Schindl, B. Schuller, B. Färber, S. Mayer, and B. Trefflich, "Online driver distraction detection using long short-term memory," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 12, no. 2, pp. 574–582, 2011.

[25] Y. Liang, J. Lee, and M. Reyes, "Nonintrusive detection of driver cognitive distraction in real time using bayesian networks," *Transportation Research Record: J. of the Transportation Research Board*, no. 2018, pp. 1–8, 2007.

[26] P. Jimenez, L. M. Bergasa, J. Nuevo, N. Hernandez, and I. G. Daza, "Gaze fixation system for the evaluation of driver distractions induced by ivis," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 13, no. 3, pp. 1167–1178, 2012.

[27] T. DOrazio, M. Leo, C. Guaragnella, and A. Distante, "A visual approach for driver inattention detection," *Pattern Recognition*, vol. 40, no. 8, pp. 2341–2355, 2007.

[28] M. Kutila, M. Jokela, T. Mäkinen, J. Viitanen, G. Markkula, and T. Victor, "Driver cognitive distraction detection: feature estimation and implementation," *Proc. of the Institution of Mechanical Engineers, Part D: J. of Automobile Engineering*, vol. 221, no. 9, pp. 1027–1040, 2007.

[29] M. E. Jabon, J. N. Bailenson, E. Pontikakis, L. Takayama, and C. Nass, "Facial expression analysis for predicting unsafe driving behavior," *IEEE Pervasive Computing*, no. 4, pp. 84–95, 2010.

[30] A. Ragab, C. Craye, M. S. Kamel, and F. Karray, "A visual-based driver distraction recognition and detection using random forest," in *Image Analysis and Recognition*. Springer, 2014, pp. 256–265.

[31] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *Intelligent Transportation Systems, IEEE Trans. on*, vol. 16, no. 1, pp. 51–65, 2015.

[32] H. Gao, A. Yuce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Image Processing (ICIP), 2014 IEEE Int. Conf. on*. IEEE, 2014, pp. 5961–5965.

[33] S. Mattes, "The lane-change-task as a tool for driver distraction evaluation," *Quality of work and products in enterprises of the future*, pp. 57–60, 2003.

[34] *ISO 26022:2010:: Road vehicles. Ergonomic aspects of transport information and control systems. Simulated lane change test to assess in-vehicle secondary task demand*, ISO Std.

[35] J. L. Harbluk, P. C. Burns, M. Lochner, and P. L. Trbovich, "Using the lane-change test (lct) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces," in *Proc. of the 4th Int. driving symposium on human factors in driver assessment, training, and vehicle design*, 2007, pp. 16–22.

[36] J. Engström and G. Markkula, "Effects of visual and cognitive distraction on lane change test performance," in *Proc. of the 4th Int. driving symposium on human factors in driver assessment, training, and vehicle design*, 2007, pp. 199–205.

[37] M. L. Turner and R. W. Engle, "Is working memory capacity task dependent?" *J. of memory and language*, vol. 28, no. 2, pp. 127–154, 1989.

[38] M. Chan and A. Singhal, "The emotional side of cognitive distraction: Implications for road safety," *Accident Analysis & Prevention*, vol. 50, pp. 147–154, 2013.

[39] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.

[40] H. Gao, H. K. Ekenel, and R. Stiefelhagen, "Pose normalization for local appearance-based face recognition," in *Advances in Biometrics*. Springer, 2009, pp. 32–41.

[41] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Computer Vision (ICCV), 2011 IEEE Int. Conf. on*. IEEE, 2011, pp. 937–944.

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conf. on*. IEEE, 2014, pp. 1701–1708.

[43] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: a 3d facial expression database for visual computing," *Visualization*

and Computer Graphics, IEEE Trans. on, vol. 20, no. 3, pp. 413–425, 2014.

[44] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conf. on*. IEEE, 2013, pp. 532–539.

[45] N. Faggian, A. Paplinski, and J. Sherrah, "3d morphable model fitting from multiple views," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE Int. Conf. on*. IEEE, 2008, pp. 1–6.

[46] C. Ecabert, H. Gao, and J.-P. Thiran, "3d bilinear face model fitting from multiple cameras," EPFL, Tech. Rep., 2015 - (Text provided as supp. mat.).

[47] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *Proc. of Automatic Face and Gesture Recognitions (FG 2015), 11th IEEE Int. Conf. on*, 2015.

[48] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[49] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Trans. on*, vol. 3, no. 1, pp. 5–17, 2012.

[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[51] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10, pp. 1771–1787, 2008.

[52] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using sift features and mean shift," *Computer vision and image understanding*, vol. 113, no. 3, pp. 345–352, 2009.

[53] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang, "Facial action unit event detection by cascade of tasks," in *Computer Vision (ICCV), 2013 IEEE Int. Conf. on*. IEEE, 2013, pp. 2400–2407.

[54] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Lett.*, 2014.

[55] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Comput. Soc. Conf. on*. IEEE, 2010, pp. 94–101.

[56] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE Int. Conf. on*. IEEE, 2011, pp. 921–926.

[57] J. M. Girard, J. F. Cohn, and F. De la Torre, "Estimating smile intensity: A better way," *Pattern Recognition Lett.*, 2014.

[58] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, vol. 92, p. 120, 2001.

[59] E. Kroupi, J.-M. Vesin, and T. Ebrahimi, "Phase-amplitude coupling between eeg and eda while experiencing multimedia content," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conf. on*. IEEE, 2013, pp. 865–870.

[60] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. of the 4th Int. Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 65–72.

[61] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc. of the 3rd ACM Int. workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[62] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proc. of the 3rd ACM Int. workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[63] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. of the 4th Int. Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

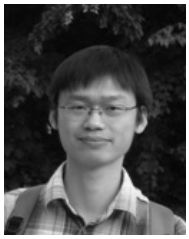
[64] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[65] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The J. of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[67] S. Boucenna, A. Narzisi, E. Tilmont, F. Muratori, G. Pioggia, D. Cohen, and M. Chetouani, "Interactive technologies for autistic children: a review," *Cognitive Computation*, vol. 6, no. 4, pp. 722–740, 2014.

Anıl Yüce received his B.Sc. from Middle East Technical University, Turkey in 2008 and M.Sc. from Ecole Polytechnique Fédérale de Lausanne, Switzerland (EPFL) in 2010 in electrical engineering. Since then he is pursuing a PhD degree at the Signal Processing Laboratory (LTS5) at EPFL. His main research interest is facial image analysis for various applications, particularly analysis of facial expressions and their dynamics. He is a student member of the IEEE since 2011.



Hua Gao received the Dipl.-Inf. and Ph.D. degrees in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2008 and 2013, respectively.

He is currently a post-doc at EPFL, Lausanne, Switzerland. His research interests include the fields in facial image processing, e.g. face tracking, 3D face reconstruction, facial expression recognition and face recognition.

Gabriel L. Cuendet received his B.Sc. and M.Sc. degrees in electrical engineering with specialization in biomedical engineering from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 2012, where he is currently working toward the Ph.D. degree in developing facial image analysis for medical diagnosis applications. The research is focused on computer vision methods for 2D and 3D facial landmarks detection and tracking.

Jean-Philippe Thiran is Associate Professor of Image Processing and director of the Signal Processing Laboratory (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an Associate Professor position with the Department of Radiology of the University Hospital Center (CHUV) and University of Lausanne (UNIL). His research field is image analysis and multimodal signal/image processing, with applications in many domains including medical image analysis, human-computer interaction, remote sensing of the Earth, and surveillance. Dr Thiran is author of co-author of more than 130 journal papers, 9 book chapters, more than 185 papers in peer-reviewed proceedings of international conferences, and holds 4 international patents. He is currently an associate editor of the IEEE Transactions on Image Processing and a reviewer for many journals and conferences. He is a senior member of the IEEE.