Activity and Stress Estimation Based on *OpenPose* and Electrocardiogram for User-Focused Level-4-Vehicles

Fabian Walocha, Uwe Drewitz, and Klas Ihme

***Abstract*—Increasing vehicle automation changes the role of humans in the car, which imposes new requirements on the design of in-vehicle software and hardware for flexible interior concepts. An option to meet these requirements is the development of user-focused automation based on combined user and context monitoring in real time. The system behavior may be dynamically adapted by adjusting the driving style or the interior lighting. Here, we present a hierarchical approach on the basis of semantically motivated low-level features for activity and stress recognition based on OpenPose and electrocardiogram data. A driving simulator study with 29 participants was conducted to determine the potential of the approach. Participants had to accomplish different tasks: manual driving (MD); mobile office work with varying task load levels (high task load: MO-HT, low task load: MO-LT); and relaxing (REL) during automated driving. The validation revealed that our model is able to correctly distinguish between different activities using only a set of primitive features (average precision: driving: 76% and mobile office work: 93%, relaxing: 86%). Furthermore, we evaluated a person-independent and a person-specific approach for stress detection and found that both strategies show similar trends in accordance with our predictions (person-independent: stress detected in MO-HT: 22%, MO-LT: 18%, MD: 18%, REL: 15%; person-specific: stress detected in MO-HT: 79%, MO-LT: 72%, MD: 65%, and REL: 50%). These results demonstrate the efficacy of using a lightweight semantic approach for activity recognition and stress detection as basis for user-focused vehicle automation.**

***Index Terms*—Activity recognition, driver monitoring, mobile office, stress estimation, user-focused automation, vehicle automation.**

I. Introduction

A. Promises and Challenges of Automated Driving

INCREASING vehicle automation will change the role of humans in the car. While nowadays humans are in charge of the driving task, soon humans will be expected to monitor the automation and to be the fall back at system boundaries. In highest automation levels (levels 4 and 5 according to SAE international, see [1]), humans are foreseen to be relieved from control at least for a certain road section. They can then devote their time and attention to activities such as relaxing, reading, or even using their car as a mobile office (e.g., [2]–[4]). Yet, users of automated vehicles have different needs than drivers of manually driven cars, such that that changing use cases also impose new requirements on the design of in-vehicle software and hardware. For instance, persons who want to prepare a presentation for a meeting at the destination need to be sure that the time is sufficient to complete the task. They do not want to resume control too early (which would be possible in SAE level 4 at section boundaries, such as the transition from highway to rural roads). To add, people may suffer from kinetosis while working on a laptop in the automated vehicle [5]. Notably, mobile office workers with high task load may want to have different information or configurations of the vehicle interior than users who do routine tasks, want to relax, or manually drive, so that vehicles that offer multiple automation levels will have to be able to adapt to the user in the current situation. An option to realize this is the development of user-focused automation which places two basic human needs at the center of system design [6]: the need to understand and the need to be understood.

The need to understand is required for goal-oriented interaction with the environment and enables understanding and predictability. To address this need, automated systems must behave in a predictable manner, so that the systems are transparent to their users [6]. The need to be understood is necessary to build a relationship and thus lays the foundation for trust and the experience of positive emotions. Automated vehicles need to be able to infer when users are uncertain or stressed and recognize when it is appropriate to provide information. To realize user-focused automation, systems have to focus on the human being by combining user and context monitoring using various sensors in real time [6]. Then, the system behavior could be adapted by adjusting, for example, the driving style, the information provided via a human-machine interface, or the interior lighting ([4], [6]). As a prerequisite for adapting towards the user, user-focused systems require a reliable real-time estimate of what the user is doing and how the user is feeling. Hence, the goal of this article is to develop a user model that is able to estimate a person's activity and state (i.e., manual driving, relaxing, and mobile office work) in real time based on video recordings and integrate this with a heart-rate-based stress assessment as the

Manuscript received October 18, 2020; revised May 11, 2021, December 9, 2021, and January 29, 2022; accepted January 31, 2022. This work was supported by the German Federal Ministry of Transport and Digital Infrastructure in the funding program Automated and Connected Driving for the Project AutoAkzept (FKZ: 16AVF2126A). This article was recommended by Associate Editor T. H. Falk. *(Corresponding author: Klas Ihme.)*

The authors are with the Institute of Transportation Systems, German Aerospace Center, D-38108 Braunschweig, Germany (e-mail: fabian.walocha@dlr.de; uwe.drewitz@dlr.de; klas.ihme@dlr.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3155375.

Digital Object Identifier 10.1109/THMS.2022.3155375

basis for satisfying the need to be understood in user-focused level-4-vehicles.

B. Human Sensing in Autonomous Driving

In artificial intelligence, human sensing is the automated perception of particularities in the state of a human. This state can be their action state (i.e., the activity the human is conducting), their physiological or mental state (i.e., the emotion and cognitive activity of the person), as well as their social state (i.e., their standing in their social group and society in general) in relation to the overall context. Next to assessing the presence of vulnerable road users outside the vehicle, driver monitoring is the most relevant use case for human sensing in the context of driving. For instance, monitoring the driver through various modalities is used to measure the capability of drivers to operate the vehicles directly by detecting whether their gaze is oriented towards the road [7]. More indirectly, these modalities can be used to identify drivers' physiological or mental state, for example, their drowsiness (e.g., [8]) or workload (e.g., [9]). In automated vehicles, a further possible application of human sensing is the detection and tracking of human activity.

A challenge for detecting and tracking human activities in the wild from raw video data lies in the noisiness and variability of the video signal. This can be due to resolution, varying lighting conditions, and, on the human side, general variability in human body height, color of skin, and the color and style of clothing. Recent advances in body pose estimations using convolutional neural networks enable the accurate detection of key joints of the human body for driving scenarios. The subfield of deriving human activity from key joints is referred to as skeleton-based action recognition [10]. By focusing on the location of the joints in the user, the challenge in activity recognition is shifted from image processing to structuring the data for classification with the advantage of heavily reducing the complexity of the incoming data.

For activity recognition, classification is traditionally done using hidden Markov models, decision trees, or support vector classifiers (SVCs) ([11], [12]). Recently, end-to-end learning using three-dimensional (3-D) convolutional networks or recurrent neural networks gained attraction in employing temporal sensitive activity recognition over a latent representation space [13]. While end-to-end solutions generally achieve better performance over large and inhomogeneous sets of data, it is a challenge to interpret the results and trace back how they are derived. In comparison, hierarchical semantically driven models are used in order to buffer the model using low-level feature matrices on which higher level reasoning is performed. That way, given the final result of hierarchical algorithms, we are able to effectively estimate the state and influence these primitive features on the classification result. Specifically, these low level features can hold information on key actors of the human body which influences a given activity [14]. Hence, these hierarchical approaches increase the interpretability of models for activity recognition.

As opposed to physical activity, mental states cannot be directly assessed but only inferred via self-report or by analyzing behavioral or physiological cues known to correlate with given mental states. For example, when humans are afraid, generally, their heart rate increases, they begin to sweat more and to breath faster [15]. Similarly, stress can be assessed based on physiological signals. Stress can be seen as a functional response of the human to environmental challenges. Stress mostly comes along with elevated arousal because the organism (i.e., brain and body) is in a state of preparedness to the requirements of the stressor [16]. Hence, in this article, stress is understood as a long-lasting state of continuous mental arousal resulting from specific task requirements. Stress has previously been found to have negative effects on memory, concentration, and overall health [17]. Especially in the context of driving, concentration is required to operate the vehicle safely. Yet, in high automation levels, the problem sphere shifts from a safety issue to comfort concerns. A user-focused system that detects if the user of an automated vehicle is stressed during his mobile office work, can be of service by removing distractors in the environment by adding focus light, optimizing the route, or turning down the volume of the radio. Stress can be assessed based on physiological measurements, such as heart rate, skin conductance, or hormonal activity, because stress affects humans' physiology and experience. In driving scenarios, drivers' stress level have been assessed using heart-rate-based measures, galvanic skin responses [18], or using biomarkers, such as salivary amylase activity [19]. However, heart-rate-based measures have the clear advantage of being robust even during movement. Moreover, next to traditional methods such as the electrocardiogram (ECG) using electrodes on the chest, flexible and/or wearable solutions for in-vehicle heart rate assessment have been developed and proven to be feasible [20], [21].

C. Goals of the Research & Contribution

In this article, we present a hierarchical, modular approach to activity recognition on the basis of semantically motivated low-level features. Our algorithm uses a sliding-window-based approach to distinguish between the activities *relaxing*, *mobile office work,* and *driving* in 60 s intervals. The activity recognition will be combined with a stress detection algorithm, in which stress is inferred by integrating a momentary assessment of the user's arousal based on heart rate (variability) over a time period of 60 s. The approach will be evaluated using realistic data from a driving simulator study with 29 participants. In the next section, we will present the model architecture. Then we will dwell on the procedure and the results of the evaluation study. Thereafter, the findings of the article are critically discussed with respect to previous literature.

II. Model Architecture

A. General Description

The proposed model is generally split into three functional units, starting with a preprocessing pipeline for video and ECG data during which the incoming data is cleaned, scaled, and augmented. The second unit contains the feature space transformation and dimensionality reduction to put the data

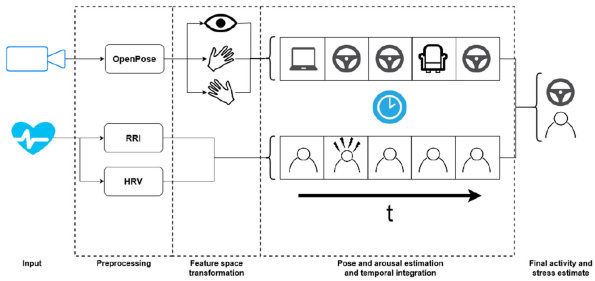

Fig. 1. Sketch of the architecture. RRI and HRV are used to detect arousal while feature primitives using hand positions and head movements determine the current pose. Final estimates are given through majority vote over a fixed time window.

into a representation space which facilitates classification. This includes the derivation of the interval between two R waves, the RR interval (RRI) in the ECG and the calculation of heart rate variability (HRV) as basis for arousal estimation and stress classification. In addition, for activity recognition, this entails deriving the relative position of the left and right hand as well as the head turning of the user from the video recordings based on OpenPose [22], yielding a set of predefined primitive features on which classification is done. This set of primitives is used in the third functional unit to classify the current pose of the user with a linear SVC. The poses, in turn, are integrated over time to yield an estimate of the participant's current activity. Additionally, we estimate instantaneous arousal using the participant's instantaneous RRI and HRV calculated over a window of the previous 60 s. The arousal value is integrated over a period of 60 s to determine whether the participant is currently in the state of stress.

We chose this hierarchical approach as opposed to an end-to-end data driven model for activity and stress detection to achieve the following benefits (a sketch of the architecture is found in Fig. 1).

1) *Sensor Agnosticity*: By employing a low-level feature buffer, we separate the activity state classification from the sensor array it is derived from. This means that the classifier can still be used in a different sensor setup as long as the low-level features can still be derived.
2) *State Agnosticity*: The feature buffer also allows modification to the activity states to be recognized, as long as the states can be readily estimated from the low-level features.
3) *Scenario Agnosticity*: The hierarchical model allows for generalizability irrespective of the scenario and context that the data is recorded.
4) *Explainability and Adaptability*: As the goal of this research is to ultimately derive adaptation strategies based on the user state, we expect the low-level features to be meaningful to select the most optimized strategy as they themselves already hold semantic information about the current user state.

B. Pose and Activity Recognition

1) Preprocessing: We use video data of the user sitting in the driver seat as the input signal. The raw video data is processed using the *OpenPose* single person model which extracts key body parts as (x and y-) coordinates in the upper body. We scale the coordinates to be between (00) and (11) based on the position of the participant's right ear as upper left corner to achieve height invariance among participants. Furthermore, we derive the position of the participant's fingertips from the position of the coordinates of the wrists and elbows provided by OpenPose by extending the vector by a factor of 1/(1.6) as we hold the fingertip position to be more meaningful as the wrist position for the respective activities, especially when working on the computer. We use a factor of 1:1.6 between hand and forearm length reflecting the classical view on ideal body proportions approaching the golden ratio [23].

2) Feature Space Transformation: The instantaneous pose is estimated based on a set of predefined, meaningful primitive features. Since the aim is to classify based on the functional component the user is exerting instead of raw image coordinates, features are extracted by determining the spatial distance of the hand coordinates toward a set of manually selected functional regions of interest (ROIs), including the steering wheel, the keyboard, the mousepad, the user's lap, and the user's head. Additionally, we approximate where a user is looking by clustering head rotation patterns.

Spatial distance of the hands towards the set of ROIs is quantified probabilistically using Gaussian mixture modeling [24]. For this, we fix the means of 2-D Gaussian distributions on the image space to the positions corresponding with these ROIs. Correlation matrices for Gaussian distributions are estimated automatically using the expectation maximization algorithm [25].

Head rotation patterns are obtained using clustering on key joints pertaining to the head (i.e., the ears, eyes, nose, and neck). This is done by first reducing dimensionality of the x and y coordinates of each of these joints using principal component analysis (PCA, [26]). The PCA transforms vector spaces onto the directions of the principal eigenvectors and is normalized based on the explained variance in each direction, yielding a uniform extension of data points in the new space. Additionally, we reduce the dimensionality by projecting the joint coordinates onto the axes which collectively account for 95% of the expected variance, effectively reducing noise and low-information movement from our data space. Moreover, we remove the computational complexity of our subsequent sampling through the dimensionality reduction. Since our new feature space is uniformly expanding, we cluster the vector space using K-means clustering [27] with Euclidean distance as our centroid metric. In order to determine the optimal number of clusters (i.e., the head positions which are meaningful and best separable given our input data), we compare clustering results using varying amounts of clusters and the resulting clusters using Davies–Bouldin-index [28]. The Davies–Bouldin-index is an indicator of cluster separation given by the ratio of within-cluster distances over between-cluster distances. Thus, compact clusters, which are farther apart from each other, are preferred. Lastly, the feature space is sparsified, yielding only the best cluster association (together with a confidence estimate) to remove positional information from the subsequent pose classification.

By removing positional information, the model is easier adapted to different setups and configurations.

3) Pose Estimation and Activity Recognition: Classification of the instantaneous pose over the set of primitive features is done using a linear SVC (SVC, [29]) given our ground truth labels data for poses *driving, mobile office work,* and *relaxing* in the training dataset. SVCs estimate hyperplanes which optimally separate labeled data by not only minimizing the error on the training data but by additionally maximizing the margin between the hyperplane and close support vectors given a regularization parameter C, which determines the relative importance of wide margins and penalty of misclassified points. Linear SVCs generalize well on new data due to the optimal margin property. Additionally, predicting a new sample given the hyperplane is fast, maintaining online capability. After the instantaneous pose is estimated, the activity is inferred as the most frequent pose using a sliding window approach over time period of 60 s.

C. Arousal and Stress Detection

Here, we distinguish between two states of instantaneous arousal, a low arousal state where a person experiences low stimulation and a high arousal state where stimulation is high. The arousal values are integrated over time to estimate the user's stress level. Stress is inferred over a time period, in which a lot of high arousal is detected.

Arousal is quantified using the interval between two subsequent R-peaks in the QRS interval, the RRI. From this, HRV is calculated using the root mean square of the successive differences between two *R* waves (RMSSD) [30].

According to previous research, decreased RRI (which corresponds to a faster heart rate) and lower HRV are linked to increased arousal due to increased sympathetic activation [31]. We here hold these findings as our ground truth assumption on a person's arousal state. As with pose and activity, we integrate the instantaneous arousal state over time to distinguish whether a user is in a state of stress or not.

We propose two approaches for the detection of arousal states. The first approach is person-independent and can be used when no user-dependent training is possible. The second approach is person-specific and uses Gaussian mixture models to estimate the relative distributions of high arousal states and low arousal states given a person's RRIs and HRV data collected over a sufficient time frame. Person-specific heart rate analysis is preferable when the respective baseline data are available since both, resting heart rate, and HRV, are affected by numerous person-specific factors, such as overall health, the amount of physical exercise, height, weight, and diet. Thus person-specific baselines are generally more meaningful than population wide estimations.

1) Preprocessing: Data preprocessing includes removing movement-related artifacts in the RRIs and calculating an estimate for HRV. Movement artifacts are expressed as a sudden change in the noted RRIs resulting from the detection of spurious heartbeats. Movement artifacts are thus identified and removed [see (1)]. For this, we followed the approach described in [32] by removing a RRI when the difference of two consecutive RRIs is greater than corresponding to 30 beats per minute

$$\mathrm{RRI}_j = \begin{cases} \mathrm{RRI}_j \; if \; \frac{60000}{\mathrm{RRI}_j} - \frac{60000}{\mathrm{RRI}_{j-1}} < 30 \\ \mathrm{NaN} \qquad\qquad \text{otherwise} \end{cases}. \tag{1}$$

Finally, RMSSD is calculated on overlapping windows in an interval of 60 s using

$$\mathrm{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}\left(\mathrm{RRI}_{i+1} - \mathrm{RRI}_i\right)^2} \tag{2}$$

where RRI_i is the RRI (i.e., the time interval between the ith and $(i+1)$-st *R* peak) and N the total number of *R* peaks in the interval. The window length of 60 s was chosen to allow a quick arousal estimation enabling relatively short-term adaptations to the final stress detection and is in line with recent recommendations on window length for RMSSD calculation [31], [33].

2) Arousal Estimation and Temporal Integration: The person-independent approach for arousal estimation quantifies arousal by taking into consideration population wide averages on the average resting RRI given the person's age and gender group (see [34]). If either information on the user is not available, a global average for that specific section is taken. If the RRI of users is in the upper 95th percentile in their category (meaning their RRI is very unlikely to be their resting RRI), users are detected to be in a state of high arousal. In the context of driver monitoring, this approach can be used, when a new user enters the car (i.e., before user-specific training is done) or when user-specific training is not available. For this approach, we only make use of the RRI and not HRV.

The second approach utilizes user-specific modeling in order to improve the sensitivity of the model to the user's specific physiology given our ground truth assumptions. Since arousal is linked to a decrease in the RRI (= an increase in heart rate) and a decrease in HRV, a 2-D bimodal Gaussian mixture model is fitted to the data of each user (consisting of the instantaneous RRIs and the corresponding HRV calculated over the previous 60 s, see above). High arousal is assigned to the distribution model with the lower average RRI and the lower average HRV (and vice versa for low arousal). We expect the values to be normally distributed inside these two distributions over the training period. The model returns the label of the arousal state which has the highest likelihood given the estimated distributions.

As mentioned above, stress is seen here as the experience of high arousal over a longer period of time. Therefore, *stress* is subsequently inferred analogously to our activity recognition strategy as the most frequent arousal state using a sliding window approach over 60 s windows which is updated with each incoming RRI.

III. Model Validation

A. Validation Dataset

In order to validate our model, a driving simulator study was conducted to determine the potential of the approach. A driving simulator study was chosen in order to validate in a realistic setting, where participants were filmed in dynamic lighting conditions while accomplishing different tasks. In total, participants

Fig. 2. Camera setup and typical pose during each scenario. From left to right. (a) Manual driving. (b) Relaxing. (c) Mobile office work.

had to accomplish three different tasks: manual driving, mobile office work, and relaxing during automated driving. In addition, the task load level of the mobile office work was varied in order to induce either high or low stress.

The study included 32 healthy adult participants between the ages of 18 and 62 years. Participants declared that they had a valid drivers' license and provided written informed consent to take part in the article. In line with the institute's guidelines to minimize risks for driving simulator study, vulnerability to simulator sickness, pregnancy, and acute intake of alcohol or other drugs were exclusion criteria for the study. No further exclusion criteria or control conditions were established for study participation. Participants received 5 € per commenced half hour as reimbursement for their participation. The recordings for three of the participants were either incomplete or erroneous, so that the data of 29 participants (average age = 25.5 years, standard deviation = 8.2 years, 14 females, 15 males) were included into the data analysis.

The experiment was implemented in the virtual reality lab with 360° full view at the German Aerospace Center [35]. Participants sat in a realistic vehicle mock-up and controlled the mock-up car in the driving simulation (Virtual Test Drive, Vires Simulationstechnologie, Bad Aibling, and Germany) via a standard interface with throttle, brake pedal, steering wheel, and indicators. All drives took place on a three-lane highway. The cockpit of the vehicle mock-up was equipped with a keyboard and a mouse pad that could be folded out on the right side of the driver seat. A screen was mounted in the center console on which the content for the mobile office task was displayed. Fig. 2 shows the setup with an exemplary pose of the different activities.

During the experiments, participants were recorded using a Logitech C930-E 1080p webcam mounted on the A-pillar at the passengers' side and using a wearable standard three-lead ECG recorded with a sampling rate of 500 Hz (HealthLab by SpaceBit, Eberswalde, Germany). HeathLab includes an *R* wave detection from the raw ECG and provided the beat-to-beat RR-values

The experiment consisted of four scenarios during which participants had to accomplish different tasks. Each scenario lasted roughly 15 min and took place on the same route. Before each scenario, participants received detailed instructions about their task in the next scenario, so that they could immediately start with the task once the driving scenario was started. The scenarios started with the vehicle entering the highway. During the automated drives, the vehicle drove with a speed of about 130 km/h or less if demanded by traffic. The scenarios ended with the vehicle exiting the highway.

The scenarios were as follows.

1) *Manual Driving (MD):* Participants had to drive manually on the highway for roughly 15 min. Participants were asked to adhere to the traffic rules and avoid driving faster than 130 km/h.
2) *Relaxing (REL):* Participants were instructed that they can relax while sitting in the automated vehicle.
3) *Mobile Office With Low Task Load (MO-LT):* During automated driving, participants had to conduct a mobile office task. The task consisted of answering emails from computer-generated "co-workers" wanting to make appointments. Participants then had to schedule and manage appointments using a calendar application. All tasks had to be performed using keyboard and mouse in a virtual Mozilla Thunderbird environment displayed on the screen on the center console of the mock-up car. In this version, only few emails (two were already in the inbox and nine further were received during the drive) had to be answered during the drive.
4) *Mobile Office With High Task Load (MO-HT):* The same mobile office task had to be accomplished as during MO-LT during automated driving. In order to induce stress, in this version however, many emails (eight were already in the inbox and 16 were received during the drive) had to be answered. Additionally, here participants were asked to complete all scheduling tasks and working through all incoming emails.

In order to reduce secondary effects of position and sequence as well as carryover effects, the order of the drives was randomized using Latin squares and participants took a break after each driving scenario. The breaks between the scenarios were about three minutes. This time was needed to start the next scenario, stop, store, restart the data recording, and instruct the participants for the next scenario. In the beginning, participants conducted two training rides were to become familiar with manual and automated driving in the simulator as well as the mobile office task. In total, the experiment took approximately two hours. For analysis, we only used the physiological and OpenPose data collected while driving on the highway (entering and exiting maneuvers were excluded).

B. Evaluation Criteria

For activity recognition, we expect that the conditions MD, REL, and MO-HT hold the activity labels for *driving*, *relaxing,* and *mobile office work,* respectively. The condition MO-LT is expected to yield a mixture of labels from *relaxing* and *mobile office work,* depending on how much time participants take to work on scheduling. The main method of evaluation employed is classification accuracy, together with precision and recall for each of the tested classes. Classes are balanced since all participants taken into consideration completed every condition of the experiment and took approximately the same amount of time for each condition (deviations account for the time participants take to reach the goal in MD, however all participants reached the goal in approximately the same time). As we are classifying activity over three conditions (driving versus mobile office work versus

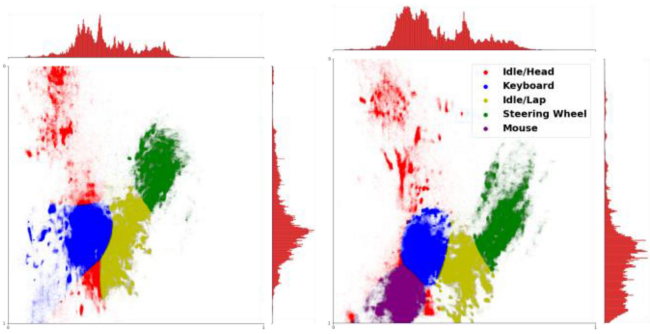

Fig. 3. Spatial cluster associations to ROIs of left and right hand respectively. The different colors refer to the different clusters the hand locations have been associated to.

relaxing) we can set the baseline (chance level) for classification accuracy to 33%.

We used a person-independent nested ten-fold cross validation in order to train and test our model. As our dataset includes 29 participants who successfully finished the experiment, we hold three persons in the testing set for each fold and use the remaining 26 participants for training the model, as well as determining the optimal penalty hyperparameter for the linear SVC. The penalty parameter yielding the highest accuracy was chosen for training via grid search over a set of candidates $C \in \{0.001, 0.10.51, 10100\}$.

As stress is not guaranteed throughout the experiment, we make the following assumptions.

1) We expect to detect the highest level of stress either in MO-HT or in MD (depending on how demanding driving is for the respective participant).
2) We expect to detect more stress in MO-HT than in MO-LT.
3) We expect to detect the least amount of stress in REL.

This was confirmed by a comparison of the average RRI of the participants between the four conditions MD (M = 837.9 ms, SD = 137.5 ms), REL (M = 862.6 ms, SD = 133.2 ms), MO-LT (M = 835.9 ms, SD = 137.4 ms), and MO-HT (M = 814.8 ms, SD = 117.2 ms) with a repeated-measures ANOVA. It revealed a main effect of condition on RRI ($F(384) = 6.9$, $p < .001$) with the following significant post-hoc comparisons: MO-HT < MO-LT; MO-HT < REL; MO-LT < REL; and MD < REL ($ps < .05$ [uncorrected], all other comparisons were not significant).

C. Evaluation Results

Given our experimental setup, assignment of each hand position to the cluster with the highest likelihood results in the clustering regions displayed in Fig. 3. Fig. 4 shows the position of the ROI centers superimposed on a heat map of hand positions. Varying the number of clusters on K-means for head rotation clustering yielded that three clusters can be best separated (lowest Davies–Bouldin score for $k = 3$: 0.7). Fig. 5 shows the points associated with each cluster. The clusters roughly correspond to participant is looking towards the center console, participant is looking straight, and participant is looking out of the left side window.

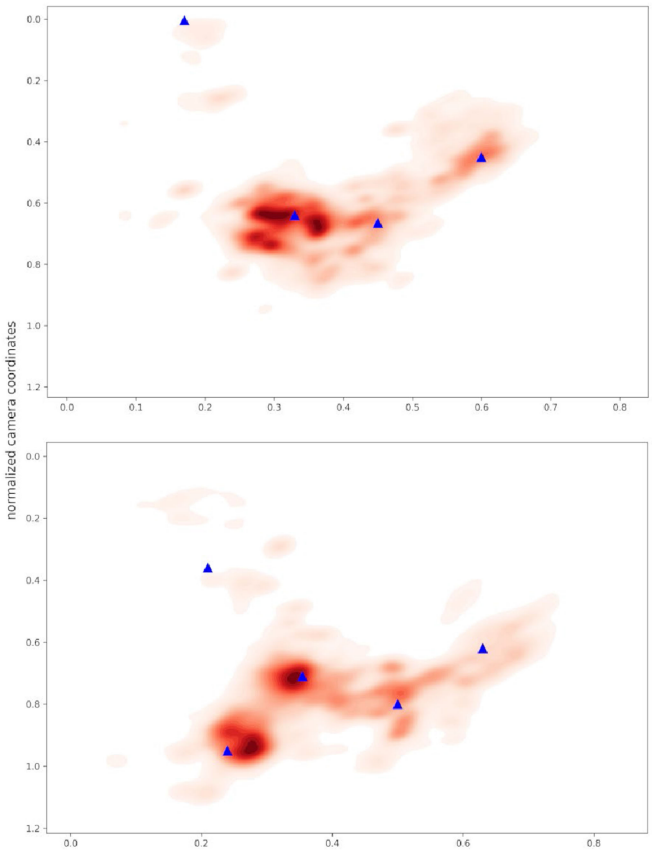

Fig. 4. Heat map of left (top) and right (bottom) hand across all conditions together with the fixed ROI centers (blue triangles).

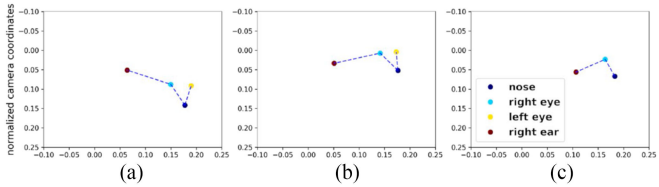

Fig. 5. Head turn cluster associations. Semantically, the three clusters refer to (a) looking towards the center console (left). (b) looking out of the front windshield (center), (c) looking out of the left-side window (right). The color of the dots denotes the respective body part (nose, right eye, left eye, right ear). Please note that (c) is missing the left eye due to occlusion during the head turn away from the camera.

Activity recognition yielded the following results: We obtained an average classification accuracy of 85% over all conditions using a linear SVC with a penalty parameter value of 0.1. Average precision is 76% for driving, 93% for mobile office work, and 86% for relaxing. The average recall over all folds is 94% for driving, 93% for mobile office work, and 74% for relaxing (given a chance level of 33%).

The label-wise classification accuracy is displayed in a confusion matrix in Fig. 6. The results show that the model had difficulties distinguishing between *driving* and *relaxing*. This can be explained by the fact that many participants choose to grab the steering wheel during driving in the low position, resulting

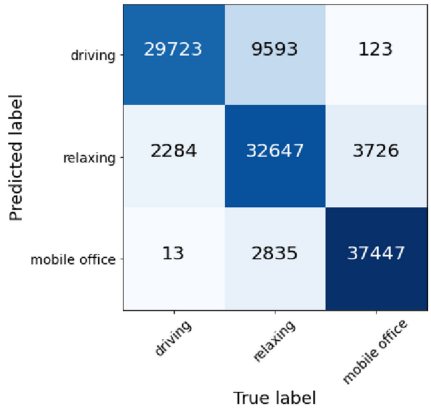

Fig. 6. Confusion matrix of SVC classifier over ten-fold cross-validation. The left shows the predicted label while the bottom shows the true label.

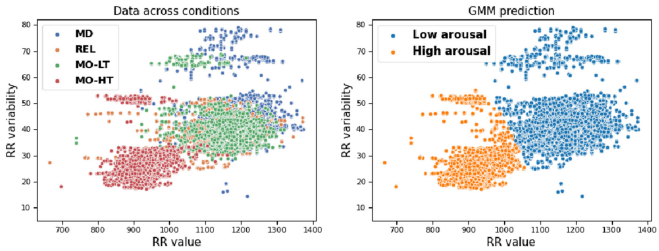

Fig. 7. RRI and HRV distribution across conditions with its corresponding arousal clustering for one participant. On the left, the association of data points to experimental conditions is displayed, while the right shows the association of data point to the arousal labels (low versus high).

in their hands being close to the lap position they hold during the relaxed condition.

Both, the person-independent threshold-based stress recognition approach as well as the person-specific GMM approach return a similar pattern, however with a different base rate. Our baseline approach detects the most stress in condition MO-HT (22% of the time) across all participants, the second most in both MO-LT and MD (18%). The least stress is detected in REL (15%).

Our person-specific mixture modeling approach detects the most stress in MO-HT (79%) across all participants, the second most stress in MO-LT (72%), third most in MD (65%), and the least stress in REL (50%). These findings are in accordance with our prediction, as MO-HT is identified as the condition which yields the highest stress and REL as the condition with the least amount of stress.

Fig. 7 shows the clustering results of arousal detection using GMM for a participant. The overall stress classification results for the two approaches in the different conditions are visualized in Fig. 8.

IV. Discussion

A. Summary of Goals and Main Results

The goal of this article was to develop a method for combined activity recognition and stress detection as building block for user-focused automation. For this endeavor, a hierarchical model

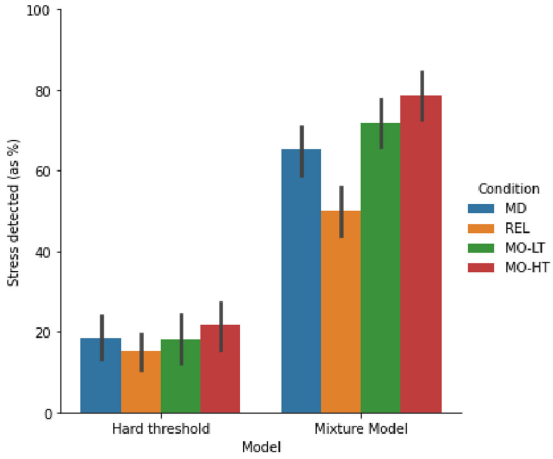

Fig. 8. Stress detection across conditions. While the baseline level detected is different between approaches, the ranking among conditions stays consistent. Error bars show the standard error of the mean.

taking video and ECG data as input, which is based on low-level semantic labels, was used. A validation on a realistic dataset from a driving simulator study revealed that our model is able to correctly distinguish between the activities driving, mobile office work, and relaxing using only our set of primitive features derived from the semantic low-level labels. Furthermore, we evaluated a person-independent and a person-specific approach for stress detection and found that both strategies showed similar trends in accordance with our prediction. These results indicate the efficacy of using a lightweight semantic approach for activity recognition and stress detection as basis for user-focused automation in the vehicle.

B. Discussion of General Architecture

The hierarchical architecture proposed in this article proved to be informative in solving the dual problem of recognizing activity and detecting stress based on a set of low-level features. By generating a semantic low-level feature buffer, we forced the model to generalize over a feature set applicable for a wide variety of scenarios. Cross-validation confirmed that the setup effectively constrains the high-level input space to a meaningful low-level embedding and still reaches good classification accuracy. While end-to-classification approaches based on deep learning can achieve relatively high classification accuracy on similar problems, such an approach on semantic low-level features has the advantage of a high transparency facilitating the interpretation of the results (see also [36]). In addition, an architecture as the one proposed here is agnostic to the specific sensors utilized to derive the low-level features and can be extended to other user states and scenarios. Worth mentioning is also the fact that deriving the low-level features was accomplished without explicit labels for them. Better labeled data may even improve the performance of the model.

C. Discussion of Activity Recognition

The approach for activity classification used in the current model provided relatively high classification accuracy for the

three different activities. In future research, this approach may be improved either by employing a wider set of static poses as primitives for the activities or by improving the recognition of our set of static poses (e.g., with a multicamera set-up). Our approach is limited because it only considers static low-level poses. This appears to be sufficient for discriminating the activities in the current dataset, but may fall short for activities that are expressed as a series of distinct poses. A recent work employed recurrent neural networks for classifying activities based on low-level primitives integrating also temporal dependencies in the activities and in this way even accomplish to discriminate activities that are more similar than the ones separated here (e.g., reading magazine versus reading newspaper or putting on jacket versus taking off jacket) [36]. Thus, our approach for activity recognition may even be improved and extended to further activities by considering temporal dependencies.

D. Discussion of Stress Detection

While both stress detection strategies revealed a similar and expected trend of the estimated stress in the different conditions, the difference in the base rate indicates that person-independent classification is challenging which is likely due to the fact that the base heart rate varies across people. Moreover, the model defines stress as elevated arousal over 60 s intervals, so that it is able to capture changes in the windows but is agnostic to changes on different time scales especially of sudden stressful events. Therefore, it may be useful to include multiple resolutions over different window sizes to capture arousal changes over different periods of time. In addition, cardiovascular indices, such as the RRI and HRV are not the only indicators for stress and especially may miss information regarding negative affect, so that the model may be improved by also considering other information, such respiration rate, facial expressions, or behavioral data to cover further aspects of stress (e.g., [37], [38]).

It also has to be mentioned that the training and validation dataset only included ground truth labels for different (high level) activities and not for different stress states, so that better labeled data with more fine-grained stress levels may improve the model performance even more. However, the analysis of the average RRIs in the four scenarios showed that the RRI decreased from REL to MO-LT and MO-HT indicating increased activation and stress. At this point, it is worth mentioning that the length of the driving scenarios was rather in the range of short drives and that stress during mobile office work may also occur during longer automated drives. To add, some stress may have resulted from the driving simulation experience instead of the accomplished activities. In addition, the sample of the evaluation study had a relatively young average age. Hence, future studies should investigate whether the developed stress estimation algorithm also works during sustained stress, in real automotive settings, and for other age cohorts.

E. Integration Into User-Focused Automation

User-focused automation [6] needs robust user modeling as input for determining the best-possible adaption strategy. Due to the wide range of user states and activities, it is unlikely that it will be possible for user-focused systems to be able to recognize all possible user states and activities. However, a robust recognition of certain states and activities may already provide the system with a possibility to adapt to these and therewith improve the interaction and experience of the user already (e.g., [6]). Hence, the model for activity and stress estimation may provide sufficient information to improve the conditions for users in dedicated mobile offices use cases [6]. Therefore, next to the abovementioned advice for improving the classification performance, an integration of the algorithm with adequate adaption strategies into first user-focused systems should be accomplished. In addition, shorter time window sizes for stress estimation may aid the design of user-focused systems for other use cases than mobile office, so that these could also react to stress elicited by sudden events (e.g., uncertainty whether the automated vehicle has correctly perceived and recognized traffic lights or vulnerable road users). In such cases, methods to detect the cardiac defense response (see [39]) may be used to supplement our approach.

V. Conclusion

Here, we presented a hierarchical model based on low-level semantic features for activity and stress recognition as initial step for the development of user-focused systems. The derived classification accuracy scores motivate to pursue this path further to realize first applications adapting to the user's current needs in (simulated) automated vehicles.

Acknowledgment

The authors thank M. Suren, A. Behrens, S. Bohmann, G. Grolms, J. Wegener, R. Möhle, and J. Rehm for their help in study preparation and data collection. In addition, we thank M. Dotzauer for proof reading.

References

[1] S. A. E. International, "Standard J3016," *Taxonomy Definitions Terms Related Road Motor Veh. Automated Driving Syst.*, vol. 4, pp. 593–598, 2014.

[2] B. Pfleging, M. Rang, and N. Broy, "Investigating user needs for non-driving-related activities during automated driving," in *Proc. 15th Int. Conf. Mobile Ubiquitous Multimedia*, 2016, pp. 91–99.

[3] K. Pollmann, O. Stefani, A. Bengsch, M. Peissner, and M. Vukelić, "How to work in the car of the future?," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–14.

[4] M. Oehl, K. Ihme, A.-A. Pape, M. Vukelić, and M. Braun, "Affective use cases for empathic vehicles in highly automated driving: Results of an expert workshop," in *Lecture Notes in Computer Science, HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design*, H. Krömker, Ed., Cham, Switzerland: Springer, 2020, pp. 89–100.

[5] C. Diels, J. E. Bos, K. Hottelart, and P. Reilhac, "Motion sickness in automated vehicles: The elephant in the room," in *Road Vehicle Automation 3*, G. Meyer and S. Beiker, Eds., Cham, Switzerland: Springer, 2016.

[6] U. Drewitz *et al.*, "Towards user-focused vehicle automation: The architectural approach of the autoakzept project," in *Lecture Notes in Computer Science, HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design*, H. Krömker, Ed., Cham, Switzerland: Springer, 2020, pp. 15–30.

[7] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? A new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intell. Transport. Syst. Mag.*, vol. 9, no. 4, pp. 10–22, Winter 2017, doi: 10.1109/MITS.2017.2743165.

[8] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors (Basel, Switzerland)*, vol. 12, no. 12, pp. 16937–16953, 2012.

[9] K. A. Brookhuis and D. de Waard, "Monitoring drivers' mental workload in driving simulators using physiological measures," *Accident; Anal. Prevention*, vol. 42, no. 3, pp. 898–903, 2010.

[10] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proc. 4th ACM Int. Workshop*, 2006, pp. 171–178.

[11] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 955–960.

[12] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Proc. 4th IEEE Int. Conf. Multimodal Interfaces*, Oct. 2002, pp. 3–8.

[13] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F.-Y. Wang, "Driver activity recognition for intelligent vehicles: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5379–5390, Jun. 2019, doi: 10.1109/TVT.2019.2908425.

[14] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver hand activity analysis in naturalistic driving studies: Challenges, algorithms, and experimental studies," *J. Electron. Imag.*, vol. 22, no. 4, 2013, Art. no. 41119, doi: 10.1117/1.JEI.22.4.041119.

[15] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[16] L. D. Sanford, D. Suchecki, and P. Meerlo, "Stress, arousal, and sleep," *Current Topics Behav. Neurosci.*, vol. 25, pp. 379–410, 2015, doi: 10.1007/7854_2014_314.

[17] A. Pietrangelo and S. Watson, The effects of stress on your body, 2018, [Online]. Available: https://www.healthline.com/health/stress/effects-on-body#1

[18] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transport. Syst.*, vol. 6, no. 2, pp. 156–166, Feb. 2005, doi: 10.1109/TITS.2005.848368.

[19] M. Deguchi, J. Wakasugi, T. Ikegami, S. Nanba, and M. Yamaguchi, "Evaluation of driver stress using motor-vehicle driving simulator," *IEEJ Trans. SM*, vol. 126, no. 8, pp. 438–444, 2006, doi: 10.1541/ieejsmas.126.438.

[20] M. Beggiato, F. Hartwich, and J. Krems, "Using smartbands, pupillometry and body motion to detect discomfort in automated driving," *Front. Human Neurosci.*, vol. 12, 2018, Art. no. 338, doi: 10.3389/fnhum.2018.00338.

[21] M. Oehler, M. Schilling, and H. D. Esperer, "Capacitive ECG system with direct access to standard leads and body surface potential mapping," *Biomedizinische Technik. Biomed. Eng.*, vol. 54, no. 6, pp. 329–335, 2009.

[22] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[23] M. Akhtaruzzaman and A. A. Shafie, "Geometrical substantiation of phi, the golden ratio and the baroque of nature, architecture, design and engineering," *Int. J. ARTS*, vol. 1, no. 1, pp. 1–22, 2012.

[24] D. N. Geary, G. J. McLachlan, and K. E. Basford, "Mixture models: Inference and applications to clustering," *J. Roy. Stat. Soc. Ser. A (Statist. Soc.)*, vol. 152, no. 1, pp. 126–127, 1989.

[25] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Jun. 1996, doi: 10.1109/79.543975.

[26] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[27] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, 1967, pp. 281–297.

[28] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[29] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[30] M. Malik *et al.*, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996.

[31] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting," *Front. Psychol.*, vol. 8, 2017, Art. no. 213, doi: 10.3389/fpsyg.2017.00213.

[32] C. Borst, W. Wieling, J. F. van Brederode, A. Hond, L. G. de Rijk, and A. J. Dunning, "Mechanisms of initial heart rate response to postural change," *Amer. J. Physiol.*, vol. 243, no. 5, pp. H676–H681, 1982, doi: 10.1152/ajpheart.1982.243.5.H676.

[33] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, 2019, Art. no. 12, doi: 10.1186/s12911-019-0742-y.

[34] I. Antelmi, *et al.*, "Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease," *Amer. J. Cardiol.*, vol. 93, no. 3, pp. 381–385, 2004.

[35] M. Fischer *et al.*, "Modular and scalable driving simulator hardware and software for the development of future driver assistence and automation systems," in *Proc. Driving Simul. Conf.*, 2014, pp. 223–229.

[36] M. Martin *et al.*, "Interior observation for cooperative vehicle handover between highly automated vehicles and drivers," (in German), *Tagungsband Der 10. VDI-Tagung Mensch-Maschine-Mobilität 2019*, Braunschweig, Germany, Springer, 2019, pp. 77–80.

[37] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J. Biomed. Inform.*, vol. 59, pp. 49–75, 2016.

[38] H. Gao, A. Yuce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 5961–5965.

[39] R. Covello, G. Fortino, R. Gravina, A. Aguilar, and J. G. Breslin, "Novel method and real-time system for detecting the cardiac defense response based on the ECG," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, May 2013, pp. 53–57.