ML-PersRef: A Machine Learning-based Personalized Multimodal Fusion Approach for Referencing Outside Objects From a Moving Vehicle

Amr Gomaa
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
amr.gomaa@dfki.de

Guillermo Reyes
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
guillermo.reyes@dfki.de

Michael Feld
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
michael.feld@dfki.de

ABSTRACT

Over the past decades, the addition of hundreds of sensors to modern vehicles has led to an exponential increase in their capabilities. This allows for novel approaches to interaction with the vehicle that go beyond traditional touch-based and voice command approaches, such as emotion recognition, head rotation, eye gaze, and pointing gestures. Although gaze and pointing gestures have been used before for referencing objects inside and outside vehicles, the multimodal interaction and fusion of these gestures have so far not been extensively studied. We propose a novel learning-based multimodal fusion approach for referencing outside-the-vehicle objects while maintaining a long driving route in a simulated environment. The proposed multimodal approaches outperform single-modality approaches in multiple aspects and conditions. Moreover, we also demonstrate possible ways to exploit behavioral differences between users when completing the referencing task to realize an adaptable personalized system for each driver. We propose a personalization technique based on the transfer-of-learning concept for exceedingly small data sizes to enhance prediction and adapt to individualistic referencing behavior. Our code is publicly available at https://github.com/amr-gomaa/ML-PersRef.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; **Pointing**; **Gestural input**; • **Computing methodologies** → *Support vector machines*; *Neural networks*.

KEYWORDS

Multimodal Fusion; Pointing; Eye Gaze; Object Referencing; Personalized Models; Machine Learning; Deep Learning

1 INTRODUCTION

As vehicles move towards fully autonomous driving, driving assistance functionalities are exponentially increasing, such as parking assistance, blind-spot detection, and cruise control. These features make the driving experience much easier and more enjoyable, which allows car manufacturers to design and implement numerous other additional features such as detecting eyes off-road using gaze monitoring, or enhancing secondary tasks while driving, such as music and infotainment control using mid-air hand gestures and voice commands. All these driving- and non-driving-related features require numerous sensors to work, such as external and internal cameras, gaze trackers, hand and finger detectors, and motion sensors. This enables researchers to design more novel features
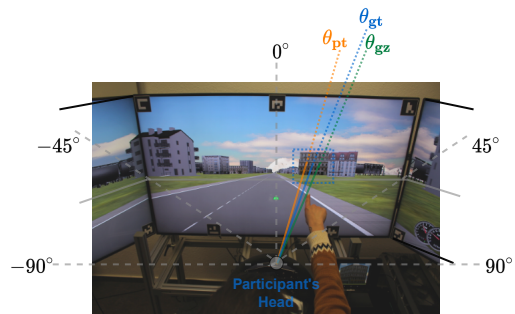

Figure 1: The Referencing Task. θ_{pt} represents the horizontal pointing angle, θ_{gz} represents the horizontal gaze angle and θ_{gt} represents the center of the referenced object.

that could make use of the full range of available sensors, such as automatic music control using emotion recognition, and referencing objects (e.g., landmarks, restaurants, and buildings) outside the car. The latter feature is the focus of this paper.

Selecting and referencing objects using speech, gaze, hand, and pointing gestures have been attempted in multiple domains and situations such as human-robot interaction, Industry 4.0, and in-vehicle and outside-the-vehicle interaction. Only recently did car manufacturers start implementing gesture-based in-vehicle object selection, such as the *BMW Natural User Interface* [1] using pointing, and outside-the-vehicle object referencing, such as the *Mercedes-Benz User Experience (MBUX) Travel Knowledge Feature* [2] using gaze. However, these features focus on single-modality interaction or limited multimodal approach where the second modality (usually speech) is used as event trigger only. This may deteriorate the referencing task performance during high-stress dynamic situations such as high-traffic driving or lane changing [8, 11, 48]. While multimodal interaction is more natural and provides superior performance compared to single-modality approaches, it introduces inherent complexity and possesses multiple challenges, such as fusion techniques (e.g., Early versus Late Fusion [9, 43, 47]), translation, co-learning, representation, and alignment [6, 7]. This has limited its use in previous systems, especially in complex dual-task situations where several modalities can be used in both tasks at the same time, such as infotainment interaction while driving.

In this paper, we explore several fusion strategies to overcome these challenges during the referencing task. Specifically, we focus on machine-learning-based approaches for personalized referencing of outside-the-vehicle objects using gaze and pointing multimodal interaction.

2 RELATED WORK

One-year-old infants learn to point at things and use hand gestures even before learning how to speak; hence, gestures are the most natural form of communication used by human beings [49]. However, gesturing is not exclusive to hand movements; it can be done using face motion or even simple gazing. Therefore, researchers have attempted various approaches for controlling objects using symbolic hand gestures [22, 38, 45, 52, 54], deictic (pointing) hand gestures [23, 25, 34], eye gaze [13, 24, 27, 31, 36, 37, 50, 51, 53], and facial expressions [3, 18, 46]. Specifically for the automotive domain, in-vehicle interaction has been attempted using hand gestures [5, 16, 33, 40], eye gaze [36], and facial expressions [44], while outside-the-vehicle interaction has been attempted using pointing gestures [17, 41], eye gaze [24], and head pose [26, 30]. Although most of the previous methods focus on single-modality approaches while using a button or voice commands as event triggers, more recent work focused on multimodal fusion approaches to enhance performance. Roider et al. [39] used a simple rule-based multimodal fusion approach to enhance pointing gesture performance for in-vehicle object selection while driving by including passive gaze tracking. They were able to enhance the overall accuracy, but only when the confidence in the pointing gesture accuracy was low. When the confidence was high, their fusion approach would have a negative effect and lead to worse results. Similarly, Aftab et al. [4] used a more sophisticated deep-learning-based multimodal fusion approach combining gaze, pointing and head pose for the same task, but with more objects to select.

While previous approaches, where both source and target are stationary, show great promise for in-vehicle interaction, they are not directly applicable to an outside-the-vehicle referencing task. This is in this case, the source and the target are in constant relative motion. Moniri et al. [32] studied multimodal interaction in referencing objects outside the vehicle for passenger seat users using eye gaze, head pose, and pointing gestures and proposed an algorithm for the selection process. Similarly, Gomaa et al. [19] investigated a multimodal referencing approach using gaze and pointing gestures while driving and highlighted personal differences in behavior among drivers. They proposed a rule-based approach for fusion that exploits the differences in users' behavior by modality switching between gaze and pointing based on a specific user or situation. While both these methods align with the multimodal interaction paradigm of this paper, they do not strictly fuse the modalities together, but rather decide on which modality to use based on performance. They also do not consider time series analysis in their approaches, but rather treat referencing as a single-frame event.

Unlike these previous methods, this paper focuses on utilizing both gaze and pointing modalities through explicit autonomous fusion methods. Our contributions can be summarized as follows: 1) we propose a machine-learning-based fusion approach and compare
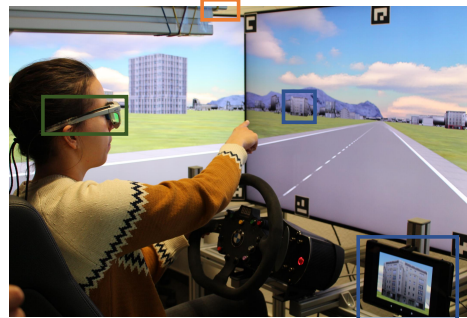

Figure 2: Our driving simulator setup overview. The pointing tracker location is highlighted in orange, gaze tracking glasses are highlighted in green, and PoI is seen in the simulator and as a notification on the tablet highlighted in blue.

it with the rule-based fusion approaches used in previous methods; 2) we compare an early fusion approach with a late fusion one; 3) we compare single frame analysis with time series analysis; 4) we utilize individuals' different referencing behavior for a personalized fusion approach that outperforms a non-personalized one.

3 DESIGN AND PARTICIPANTS

We designed a within-subject counterbalanced experiment using the same setup (as seen in Figure 2) as Gomaa et al. [19] to collect data for the referencing task. We used OpenDS software [29] with a steering wheel, pedals, and speakers for the driving simulation; a state-of-the-art non-commercial hand tracking camera prototype specially designed for in-vehicle control for pointing tracking; and SMI Eye Tracking Glasses[1] for gaze tracking. For the driving task, we used the same 40-minute five-conditions star-shaped driving route with no traffic and a maximum driving speed of 60 km/hour as Gomaa et al. [19]. The five counterbalanced conditions were divided into two levels of environment density (i.e., number of distractors around the Points of Interest (PoI)), two levels of distance from the road (i.e., near versus far PoI), and one autonomous driving condition. PoIs (e.g., Buildings) to reference are randomly located on the left and right of the road for all conditions and signaled to the driver in real-time. Drivers completed 24 referencing tasks per condition with a total of 124 gestures. These conditions were selected to implicitly increase the external validity of the experiment and the robustness of the learning models, however, there were not explicitly classified during model training. Finally, the PoIs' shapes and colors were selected during a pre-study to subtly contrast the surroundings and a pilot study was conducted to tune the design parameters.

We recruited 45 participants for the study. Six were excluded: one due to motion sickness and failure to complete the entire route, another due to incorrectly executing the task, and the other four due to technical problems with the tracking devices. The remaining 39 participants (17 males) had a mean age of 26 years (SD=6). Thirty-two participants were right-handed while the rest were left-handed.

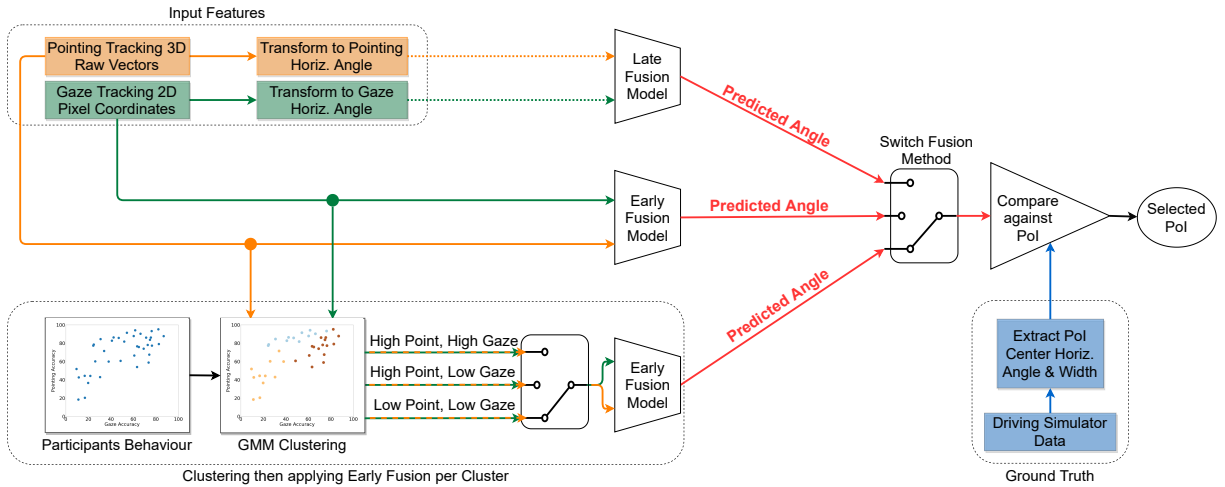[1]https://imotions.com/hardware/smi-eye-tracking-glasses/

Figure 3: System architecture showing possible fusion approaches. Early fusion is an end-to-end machine learning approach while late fusion requires pre-processing the pointing and gaze modalities first, and extracting final horizontal angles for both using the approach in [19].

4 METHOD

Figure 3 shows our proposed fusion approaches. We compared using late and early fusion on the pointing and gaze modalities to determine the referenced object. We also performed a hybrid learning approach that combines unsupervised and supervised learning to predict the PoI's center angle; we first cluster participants based on pointing and gaze behavior and then we train our early fusion model on each cluster to boost the fusion performance per user.

4.1 Fusion Approach

We consider a common 1D polar coordinate system with horizontal angular coordinates only (as seen in Figure 1) to which both pointing and gaze tracking systems as well as the PoI ground truth (GT) are transformed (similar to [19, 24]). Since this is a continuous output case, we need to solve a regression problem. We consider three fusion approaches depending on the input feature type as follows.

4.1.1 Late Fusion. The first approach involves pre-processing of the data, in which we transform the pointing and gaze raw data using the method in [19], then use these transformed horizontal angles as input features to our machine learning algorithm while using the PoI center angle as the target output (i.e., ground truth).

4.1.2 Early Fusion. The second approach is an end-to-end solution where we use the raw 3D vectors outputted from the pointing tracker and the 1D x-axis pixel coordinates outputted from the gaze tracker as input features to the learning algorithm while the ground truth remains the same.

4.1.3 Hybrid Fusion. The third approach is a hybrid approach that combines early fusion with a Gaussian Mixture Model (GMM) [21] clustering algorithm where the participants are divided into three clusters based on the accuracy of pointing and gaze (i.e., the distribution of pointing and gaze accuracy per participant as calculated

in [19]). The three clusters are "Participants with high pointing and high gaze accuracy", "Participants with high pointing and low gaze accuracy", and "Participants with low pointing and low gaze accuracy"; there were no participants with high gaze and low pointing accuracy. Then, pointing and gaze input features are considered for each cluster of participants separately when applying the early fusion method to predict the GT angle. This approach can be considered as a semi-personalized approach that exploits individual referencing performance on a group of users.

4.2 Frame Selection Method

We use the pointing modality as the main trigger for the referencing action, and thus we have multiple frame selection approaches to fuse it with the gaze modality. We define them as follows:

4.2.1 All Pointing, All Gaze. The first approach is to use all pointing and gaze frames as input modalities with the corresponding PoI center angle as GT. This produces, on average, five data points (i.e., frames) per PoI per user, resulting in an average of 620 data points per user that could be used for training or testing.

4.2.2 Mid Pointing, Mid Gaze. The second approach is to use the middle frame of pointing and the middle frame of gaze as input modalities with the corresponding PoI center angle as GT. This approach produces one data point per PoI per user, resulting in 124 data points per user that could be used for training or testing.

4.2.3 Mid Pointing, First Gaze. The third and final approach is to use the middle frame of pointing as pointing modality input with the corresponding PoI center angle as GT. However, we will take the gaze modality frame at the time of the first pointing frame. The reason behind this approach is that there is evidence that drivers tend to glance at an object only at the very beginning of the pointing action then focus on the road [19, 41]. This approach produces the same number of data points per user as the second approach.
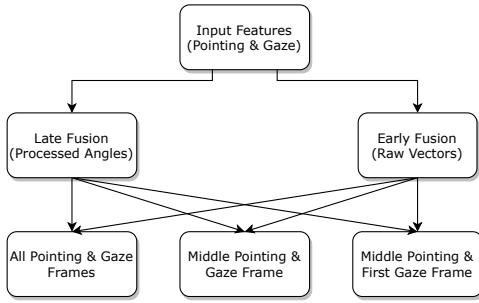
Figure 4: Input features with possible different frame selection combinations for each fusion approach.

Figure 4 summarizes these different input feature approaches along with possible frame selection methods.

4.3 Dataset Split

The entire dataset consisting of 39 participants is split into training, validation, and test sets using nested cross-validation [12] approach with 4-fold inner and outer loops. In addition, the test data is further split when training and testing personalized models (see Figure 5). Note that all the splits are done on the participant level and not on the data point level. This way, no data from the same participant is used in training and testing for the same model, to ensure external validity and model generalization. Two models are implemented to satisfy generalization and personalization aspects as follows.

4.3.1 Universal Background Model (UBM). This model follows the classic approach to machine learning: use training and validation data to create a prediction model with low generalization error reflected by its performance on the test data. A total of 29 participants are split into training and validation sets by the 4-fold inner loop of the nested cross-validation. It is then evaluated on the 10 participants selected for the test set by the 4-fold outer loop.

4.3.2 Personalized Model (PM). To train and test the personalization effect, we split the test data of each of the 10 participants into two halves, one that will be used as training data for personalization, and the other for personalization testing. Note that data points are shuffled before splitting for each participant. For each participant, we add the first half of the participant's samples to the training samples from the UBM (that corresponds to 29 participants) and we retrain a Personalized Model (PM) that was adapted for this participant. This trained model is tested on the second half of the same participant's data points as well as the data points of each of the other 9 test participants to verify the effect of personalization.

4.4 Fusion Model

For each of the fusion approaches and different frame selection approaches, we use a machine learning model for learning. Specifically, we implement a Support Vector machine for Regression model (SVR) [42]. We hypothesize that a deep learning model is not adequate in our use case due to the dataset's small size, however, we also compare different fusion models using deep learning approaches [10, 20] (e.g., Fully Connected Neural Network (FCNN),

Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)). Technical specifications for each model are as follows.

4.4.1 Machine Learning Model: SVR. Support Vector Machines (SVM) [15] are a powerful machine learning algorithm based on the maximal margin and support vector classifier, which were first intended for linear binary classification. However, later extensions to the algorithms made it possible to train SVMs for multi-class classification and regression (SVR) [42] both linearly and non-linearly. In our setup, we use the python-based library **Scikit-Learn** [35] for SVR implementation with a non-linear radial basis function (RBF) kernel and epsilon value of two, while the remaining parameters remain default values. These values for the hyperparameters were chosen using 4-fold cross-validation.

4.4.2 Deep Learning Models: FCNN, CNN, and LSTM. Deep learning or Artificial Neural Networks [10, 20] are, in many regards, more powerful learning algorithms than traditional machine learning ones. However, they require much more training data to function optimally [28]. We use the Python-based library **Keras** [14] for FCNN, CNN, and LSTM implementation. Due to the small size of the training dataset, we use a small number of layers for the neural network implementation to reduce the number of learnable parameters; the FCNN network consists of three hidden layers of size 32, 16, and 8 respectively, while non-linearity is introduced using the rectified linear unit (ReLU) activation function, and the 1D CNN network consists of one hidden convolutional layer of 64 filters and a kernel size of 3. Lastly, the LSTM network consists of one hidden layer having 50 hidden units. The output layer for all types of networks is a fully connected layer of output one and a linear activation function to work as regressors. Similar to the machine learning models, these architecture sizes (e.g., number of layers) and hyperparameters (e.g., filter and kernel sizes) were chosen using 4-fold cross-validation.

4.5 Performance Metrics

The most commonly used metric for regression problems is Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), which would directly relate to the PoI center angle (i.e., GT). However, in our scenario, users might point (and gaze) at the edge of the PoI instead of the center, which still should be considered as correct referencing. For instance, a user might point at the edge of a wide building with a higher MSE than when pointing at the edge of a tall, narrow building. Therefore, we define another metric relating to the PoI width for performance comparison along with the RMSE.

4.5.1 Root Mean Square Error (RMSE). Root Mean Square Error (RMSE) relates directly to the output angle. Figure 6 further illustrates this RMSE metric. The machine learning model attempts to predict a fused angle of pointing and gaze ($\theta_{predicted}$) as close to the PoI center (i.e., ground truth) as possible. The difference between the predicted angle and the ground truth angle (θ_{error}) is calculated for each data point and squared; then, the mean of all N square errors is calculated, which results in the MSE. Finally, the RMSE is simply the square root of the MSE, which produces an average estimation of prediction error in degrees as seen in Equation 1.
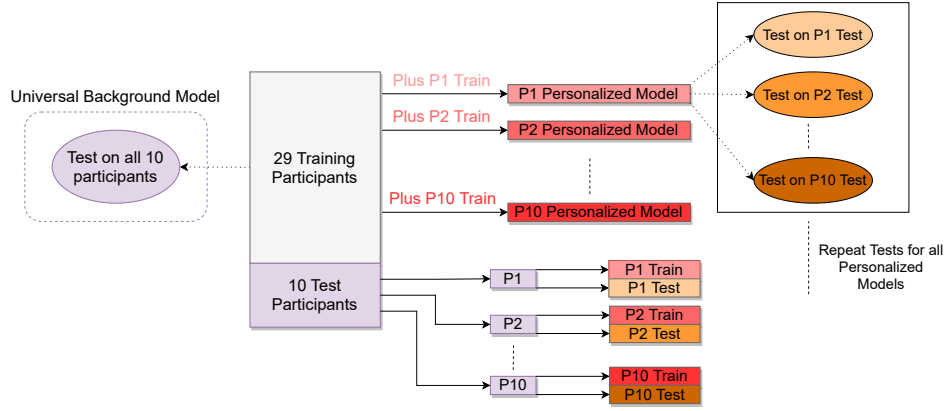
Figure 5: Dataset split approach for UBM and PM.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\theta_{error_i}^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_{prediction_i} - \theta_{GT_i})^2} \quad (1)$$

For example, if the resultant MSE of our prediction is 225, it means that the average error of the model prediction (i.e., RMSE) is 15 degrees away from the center of the PoI, so if we assume that we have a PoI's angular width of 40 degrees in visual space, it means that our model already predicts a referencing angle completely within the PoI. Note that a limitation of the RMSE metric is that it does not differentiate between positive and negative offset since it is squared during the MSE calculation. So, this 15-degree average error could be to the left or the right of the PoI center.

4.5.2 Mean Relative Distance-agnostic Error (MRDE). The previous RMSE metric returns an average absolute error in degrees with
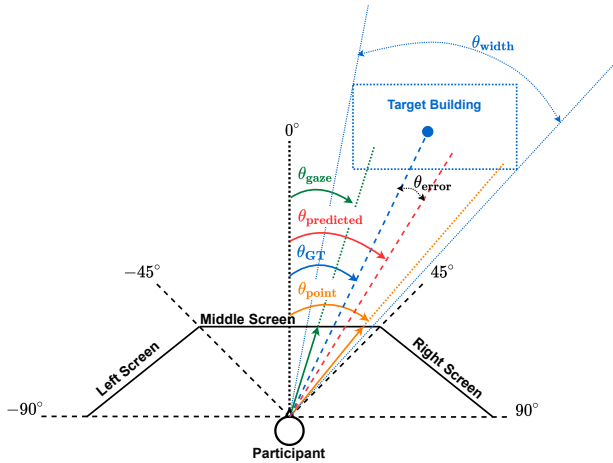

Figure 6: An illustration showing example pointing and gaze vectors and their respective angles. It also shows the predicted angle out of these vectors and how it compares to the ground truth angle.

which we, in theory, could assess our model performance. However, when interpreting this error, caution must be taken, since larger error values have a more severe impact for far PoIs versus near ones as far objects have much lower angular width (θ_{width}) compared to closer objects. Thus, a relative metric would better describe the performance in comparison to an absolute one. Consequently, we introduced the Relative Distance-agnostic Error (RDE) metric, which divides the difference between the predicted angle and the ground truth angle (θ_{error}) by half of the angular width ($\frac{\theta_{width}}{2}$) of the PoI on the screen at this time instance (since the vehicle is moving, the angular width changes with time). The resulting value then describes the error in multiples of half the PoI's width at each time instance. Thus, the MRDE is the average of the RDE for all N time instances (i.e., data points) as seen in Equation 2. Consequently, values less than one denote that the pointing vector lies within the PoI, while values larger than one denote that the referencing vector is outside the PoI's boundaries.

$$MRDE = \frac{1}{N}\sum_{i=1}^{N}\frac{\theta_{error_i}}{0.5 \cdot \theta_{width_i}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\theta_{prediction_i} - \theta_{GT_i}}{0.5 \cdot \theta_{width_i}} \quad (2)$$

While this metric allows a better interpretation of the severity of the referencing error in this study, it still does not yield a perfect absolute rating, since it now depends on the size of PoIs used in this study. For the task at hand, however, we consider this sufficient since PoI sizes did not vary greatly across the experiment.

5 RESULTS AND DISCUSSION

Results are divided into three sections. Section 5.1 shows and discusses the UBM results and the selection of the best fusion approach. We implement the UBM using all fusion approaches, frame selection methods, and fusion models to determine the optimum fusion combination that is used in the personalization approach. Section 5.2 discusses the PM results and the selection of additional hyperparameters that are concerned with personalization only. Section 5.3 discusses some further investigations on the pointing behavior quantitatively. Additionally, we illustrate the effect of multimodal fusion in comparison to single-modality approaches in all sections.

5.1 Universal Background Model (UBM) Results

In this section, we first show the results of using a machine learning model, specifically SVR. Then, we compare against the FCNN and hybrid models. Lastly, we illustrate the effect of including data points in chronological order in the analysis (i.e., time series analysis) using CNN and LSTM.

5.1.1 UBM's Early vs. Late Fusion using SVR. Figure 7a and Figure 7b show the comparison between early and late fusion approaches, using the RMSE and MRDE metrics respectively, against the single-modality approach, where lower values mean better performance as discussed earlier. Early fusion is superior to late fusion for RMSE values with a relative average of 3.01%, while MRDE values are almost the same. Moreover, this early fusion approach outperforms pointing and gaze single-modality approaches by 2.82% and 12.55% respectively for RMSE and by 5.79% and 18.55% respectively for MRDE. Pointing modality is more accurate than gaze modality by 11.02% average enhancement in RMSE and MRDE, and using raw pointing data outperforms pre-processing the pointing modality using the method in [19] by 12.01% average enhancement in RMSE and MRDE. These results agree with previous work in multimodal interaction [4, 19, 39, 40] showing that while human interaction is multimodal in general, one modality would be dominant while the other modalities would be less relevant and their effect less significant in the referencing process.

The figures also illustrate the differences between possible pointing and gaze frame selection methods: while the performance difference is not statistically significant for any frame selection method over another, selecting the middle pointing frame and the first gaze frame still outperforms the other frame selection methods by an average of 2.65% and 4.64% for RMSE and MRDE respectively. This result also agrees with previous work [19, 41] where participants usually glance at PoIs only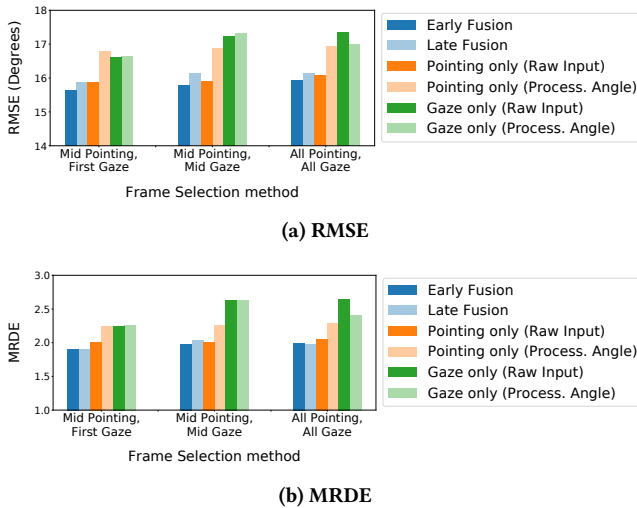 at the beginning of the referencing task, while keeping their eyes on the road and pointing at the PoI during the remaining time of the referencing action.

5.1.2 Comparison against FCNN and Hybrid Models. Having chosen the "middle pointing and first gaze" frame selection approach as an optimum approach, Figure 8a and Figure 8b show the comparison between the SVR model and the FCNN model. Early fusion using the SVR model is superior to using the FCNN model for both RMSE and MRDE values with a relative average of 2.91% and 8.75% respectively. Additionally, both pointing and gaze single-modality prediction using SVR is on average more accurate than using FCNN for both RMSE and MRDE by 4.57% and 13.14% respectively. Since the entire training dataset consists of only 3596 data points (i.e., 29 participants with 124 data points per participant), we attribute the superiority of SVR over FCNN to the small dataset size in our use case. This is because deep neural networks generally require more data (even for the shallow FCNN used), but this remains to be tested in the future. Finally, the figures show a comparison between early fusion and hybrid fusion approaches. Fusion clusters (including high pointing accuracy, high gaze accuracy, or both) outperform the "no GMM clustering" approach for RMSE and MRDE by an average of 33.71% and 3.38% respectively, while the cluster where both modalities performed poorly (i.e., low pointing and low gaze accuracy) is degraded by 49.80% and 43.45% respectively. Moreover, for all clustering and no clustering cases, fusion models achieve better performance than single-modality ones. This means that the fusion model works better than the single best modality and does not get degraded by the other modality, unlike the rule-based fusion model in [39] that degrades with high pointing accuracy.
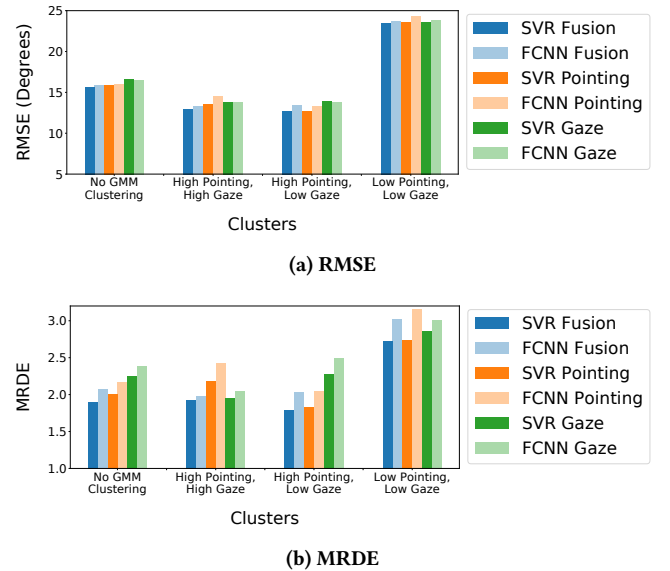

(a) RMSE

(b) MRDE

Figure 7: RMSE and MRDE results comparing UBM's Early versus Late Fusion for the multimodal interaction against prediction using single modality pointing and gaze for all frame selection approaches.

(a) RMSE

(b) MRDE

Figure 8: RMSE and MRDE results comparing UBM's Early Fusion versus Hybrid Early fusion approach when using the SVR algorithm or FCNN against single modalities. The results are calculated using the "middle pointing and first gaze" frame selection method.

5.1.3 Time Series Analysis using CNN and LSTM. Since both the CNN and LSTM algorithms require temporal information, the "all pointing and all gaze" frame selection method is used in analyses for both CNN and LSTM models as well as the previous SVR model for comparison. Figure 9a and Figure 9b show the results for these analyses while comparing the early fusion model to the hybrid fusion model as in the previous section. From the figures, it can be seen that while the temporal approach methods (i.e., CNN and LSTM) are almost equivalent for all cases in both RMSE and MRDE metrics, they are quite different from the non-temporal one (i.e., SVR). Although temporal models are slightly superior to the non-temporal model for RMSE values (Figure 9a) with an average of 3.26%, they are inferior in terms of MRDE values (Figure 9b) by an average of 7.33%, which suggests that a non-temporal approach is still the optimum approach. We hypothesize that the reason for that could also be the small dataset size, or it might be due to the fact that the temporal information for this task is too noisy or redundant to be useful for the fusion.

5.2 Personalized Model (PM) Results

As seen in Figure 5, the PM is implemented by concatenating the overall training data points (i.e., all of the 29 participants' data points) with the training data points of each of the 10 test participants separately, then retraining this new dataset with a fusion model. However, when retraining using SVR, non-support-vector data points can be disregarded, as only the support vectors returned from the UBM training contribute to the algorithm. These support vectors are usually around 600 data points, which are combined with 62 data points (i.e., half of the participants' data points) during personalization. This low ratio (around 10%) of individual

participants' data points' contribution to the retraining process significantly reduces the personalization effect. Therefore, we introduced a sampling weight hyperparameter that would increase the importance of newly added data points when retraining by putting more emphasis on these data points by the SVR regressor.

Figure 10a and Figure 10b show the RMSE and MRDE results for the PM with the effect of this sampling weight, respectively, and they compare the PM to the UBM. It can be seen that the PM is superior to the UBM even if the personalized participant training data points (i.e., personalized samples) are weighted one-to-one with the other training participants' training data points (i.e., UBM samples); however, when giving more weight to the personalized participants' samples, the model adapts to the test data of these participants until we reach a point where the model starts overfitting and it does not perform well for this participant and the others. This point can be seen around a sampling weight ratio of one-to-fifteen in both Figure 10a and Figure 10b, which shows a relative average

(a) RMSE

(b) MRDE
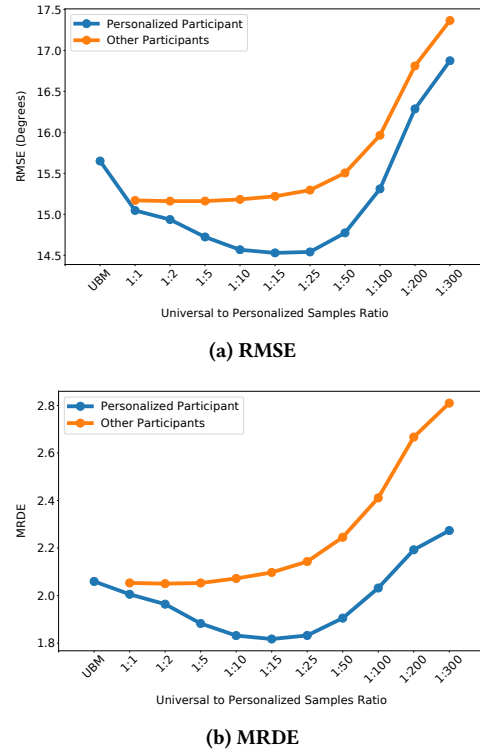
Figure 10: Fusion RMSE and MRDE results comparing PM with different sample weights when adding the personalized participant data to the trained UBM model and retraining for individual effect. The PM is also compared to the UBM model without any personalization. The "Personalized Participant" RMSE values in the graph are the average of the test results for all PMs for each participant. The "Other Participants" values are the average of the test results of the other non-personalized participants using the PM. All results are calculated using the "middle pointing and first gaze" frame selection method and the SVR algorithm.

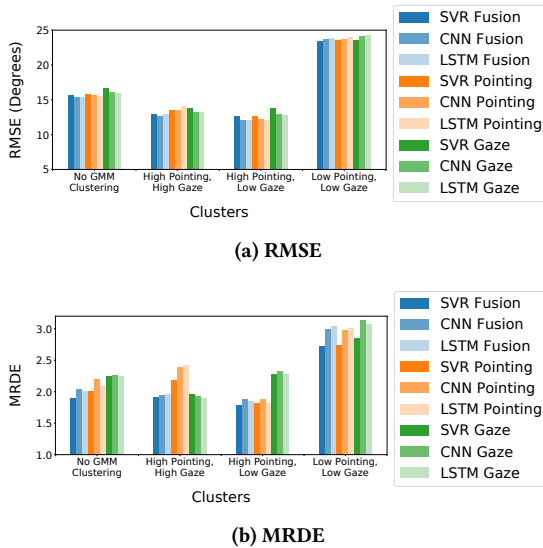

(a) RMSE

(b) MRDE

Figure 9: RMSE and MRDE results comparing UBM's Early Fusion versus the Hybrid Early Fusion approach when using the SVR algorithm, CNN or LSTM against single modalities. The results are calculated using the "middle pointing and first gaze" frame selection method.

enhancement for PM over UBM by 7.16% and 11.76% for RMSE and MRDE respectively. To further prove the personalization effect, it is tested against other participants that the model has not been personalized or adapted for. It can be seen from the same figures that, on average, PM works best for the personalized participant for all values of sample weight, with the largest gap around the one-to-fifteen optimum point. Further analysis on participants' level shows that at least 70% of the participants conform to these average figures and their referencing performance enhances with personalization. In conclusion, while there could be other personalization techniques that would enhance the referencing task performance further, these results show that our suggested PM outperforms a generalized one-model-fits-all approach.

5.3   Referencing Behavior Analysis

From Equation 2, we defined relative distance-agnostic error as the division of the difference between the predicted angle and the ground truth angle (θ_{error}) by half of the angular width ($\frac{\theta_{width}}{2}$). However, it was noticed that at the middle frame of pointing, most of the buildings have very small angular widths (θ_{width}). Additionally, it was noticed that participants' average ground truth angle (θ_{GT}) at the middle pointing frame was ±9.8 degrees (i.e., 9.8 degrees to the left or the right) and the ninety percent quantile of the ground truth angle was ±17.3 degrees. This means that most of the time participants did not wait for the building to get closer (a close building angle being around 35 to 45 degrees) and pointed rather quickly as soon as they identified it on the horizon. This is also confirmed by the average pointing time of 3.69 seconds reported in [19], which shows that the users point quite quickly when the building appears, rather than waiting, even though the building remains in their view for approximately 20 seconds.

To leverage from this observation, a slightly modified task was created by removing the pointing attempts whose ground truth exceeded ±17.3 degrees (i.e., the 90% quantile threshold) from the dataset and analyzing only pointing at faraway buildings, since pointing at close buildings corresponds to only 10% of this dataset and might not be representative. Then the PM's average RMSE and MRDE values are enhanced from 14.53 and 1.81 respectively (for the original task) to 5.45 and 0.544 respectively (for the modified far-pointing only task). This could be evidence that users attempt to point more accurately and gaze more precisely when referencing faraway objects to identify them, while they tend to be less accurate and more careless when pointing at near objects. In a real car environment, this observation is quite useful, since the surrounding world is captured using the car exterior camera and PoI angular width and distance can be easily identified. Moreover, the data was further split based on the ±9.8 average pointing angle to compare the pointing behavior at very faraway buildings (below this average angle) and faraway buildings (above this average angle). The RMSE and MRDE values for both those cases were: 4.83 and 0.64 for the very far case, and 6.67 and 0.63 for the far case. This shows a correlation between the distance of the object from the user and the referencing behavior which supports the previously mentioned hypothesis that users point more accurately at more distant buildings as seen from the better RMSE value.

6   CONCLUSION AND FUTURE WORK

Participants behave quite differently in multimodal interaction when referencing objects while driving. While previous work (such as [19]) showed pointing and gaze behavior and discussed the differences in users' pointing behavior quantitatively and qualitatively towards modality fusion, they did not investigate the implementation of the fusion model itself. In this paper, we focus more on leveraging this information and the implementation of a multimodal fusion model using both pointing and gaze to enhance the referencing task. We also compare such generalized models to a more user-specific personalized one that adapts to these differences in user behavior. Several machine learning algorithms have been trained and tested for the multimodal fusion approach and compared against the single-modality one. Also, different frame selection methods have been studied to assess the effect of time-series data on the performance when compared to single-frame methods.

UBM results show that, on average, a multimodal fusion approach is more suitable than a single modality approach. They also show that a machine learning-based approach works better than deep learning, and time series analysis does not add much enhancement in performance compared to single-frame methods. However, this could be an artifact due to the dataset's small size, which we shall address in future work with longer experiment durations to collect more data. As for the personalization effect, non-weighted PMs show a slight enhancement over the UBM, which could be attributed to the low contribution ratio of the participants' data points. On the other hand, weighted PMs show a large enhancement in the referencing task performance for that participant due to the emphasis on their data points (i.e., individual behavior).

Finally, the proposed fusion models, frame selection methods, and PM approaches enhance the referencing performance in comparison to both single-modality models and UBM. However, some aspects, like different pointing behavior based on the relative distance of the object to the user, are obscured through averaging over all data points, and this is not addressed in this study. In future work, these aspects require additional studies tailored towards this goal and further investigation to find a more concrete interpretation.

ACKNOWLEDGMENTS

REFERENCES

[1] 2019. BMW Natural Interaction unveiled at MWC 2019. https://discover.bmw.co.uk/article/bmw-natural-interaction-unveiled-at-mwc-2019
[2] 2021. Mercedes-Benz presents the MBUX Hyperscreen at CES 2021. https://media.daimler.com/marsMediaSite/instance/ko.xhtml?oid=48617114&filename=Mercedes-Benz-presents-the-MBUX-Hyperscreen-at-CES
[3] F. Abdat, C. Maaoui, and A. Pruski. 2011. Human-computer interaction using emotion recognition from facial expression. In *Proceedings of the UKSim 5th European Symposium on Computer Modeling and Simulation.* IEEE, 196–201.
[4] Abdul Rafey Aftab, Michael von der Beeck, and Michael Feld. 2020. You have a point there: object selection inside an automobile using gaze, head pose and finger pointing. In *Proceedings of the 22nd International Conference on Multimodal Interaction.* ACM, 595–603.
[5] Bashar I. Ahmad, Chrisminder Hare, Harpreet Singh, Arber Shabani, Briana Lindsay, Lee Skrypchuk, Patrick Langdon, and Simon Godsill. 2018. Selection facilitation schemes for predictive touch with mid-air pointing gestures in automotive displays. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* ACM, 21–32.

[6] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16, 6 (2010), 345–379.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[8] Shaibal Barua, Mobyen Uddin Ahmed, and Shahina Begum. 2020. Towards intelligent data analytics: A case study in driver cognitive load classification. *Brain Sciences* 10, 8 (2020), 1–19.

[9] S. Bianco, D. Mazzini, D. P. Pau, and R. Schettini. 2015. Local detectors and compact descriptors for visual search: A quantitative comparison. *Digital Signal Processing: A Review Journal* 44, 1 (9 2015), 1–13.

[10] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

[11] Eddie Brown, David R. Large, Hannah Limerick, and Gary Burnett. 2020. Ultrahapticons: "Haptifying" drivers' mental models to transform automotive mid-air haptic gesture infotainment interfaces. In *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 54–57.

[12] Michael W. Browne. 2000. Cross-validation methods. *Journal of Mathematical Psychology* 44, 1 (3 2000), 108–132.

[13] Shiwei Cheng, Jing Fan, and Anind K Dey. 2018. Smooth Gaze: A framework for recovering tasks across devices using eye tracking. *Personal and Ubiquitous Computing* 22, 3 (2018), 489–501.

[14] François Chollet and others. 2015. Keras. https://keras.io

[15] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[16] Hessam Jahani Fariman, Hasan J. Alyamani, Manolya Kavakli, and Len Hamey. 2016. Designing a user-defined gesture vocabulary for an in-vehicle climate control system. In *Proceedings of the 28th Australian Computer-Human Interaction Conference*. ACM, 391–395.

[17] Kikuo Fujimura, Lijie Xu, Cuong Tran, Rishabh Bhandari, and Victor Ng-Thow-Hing. 2013. Driver queries using wheel-constrained finger pointing and 3-D head-up display visual feedback. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 56–62.

[18] Ali Ghorbandaei Pour, Alireza Taheri, Minoo Alemi, and Ali Meghdari. 2018. Human–Robot Facial Expression Reciprocal Interaction Platform: Case Studies on Children with Autism. *International Journal of Social Robotics* 10, 2 (4 2018), 179–198.

[19] Amr Gomaa, Guillermo Reyes, Alexandra Alles, Lydia Rupp, and Michael Feld. 2020. Studying person-specific pointing and gaze behavior for multimodal referencing of outside objects from a moving vehicle. In *Proceedings of the 22nd International Conference on Multimodal Interaction*. ACM, 501–509.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[21] D. J. Hand, G. J. McLachlan, and K. E. Basford. 1989. Mixture models: Inference and applications to clustering. *Applied Statistics* 38, 2 (1989), 384–385.

[22] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi S. Nayak. 2017. Hand gesture recognition for human computer interaction. *Procedia Computer Science* 115 (2017), 367–374.

[23] Pan Jing and Guan Ye-Peng. 2013. Human-computer interaction using pointing gesture based on an adaptive virtual touch screen. *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6, 4 (2013), 81–91.

[24] Shinjae Kang, Byungjo Kim, Sangrok Han, and Hyogon Kim. 2015. Do you see what I see: Towards a gaze-based surroundings query processing system. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 93–100.

[25] Roland Kehl and Luc Van Gool. 2004. Real-time pointing gesture recognition for an immersive environment. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*. IEEE, 577–582.

[26] Young-Ho Kim and Teruhisa Misu. 2014. Identification of the driver's interest point using a head pose trajectory for situated dialog systems. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 92–95.

[27] Michael Land and Benjamin Tatler. 2009. *Looking and acting: Vision and eye movements in natural behaviour*. Oxford University Press.

[28] Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

[29] Rafael Math, Angela Mahr, Mohammad M Moniri, and Christian Müller. 2013. OpenDS: A new open-source driving simulator for research. *GMM-Fachbericht-AmE 2013* (2013).

[30] Teruhisa Misu, Antoine Raux, Rakesh Gupta, and Ian Lane. 2014. Situated language understanding at 25 miles per hour. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL, 22–31.

[31] Mohammad Mehdi Moniri. 2018. *Gaze and Peripheral Vision Analysis for Human-Environment Interaction: Applications in Automotive and Mixed-Reality Scenarios*. Ph.D. Dissertation.

[32] Mohammad Mehdi Moniri and Christian Müller. 2012. Multimodal reference resolution for mobile spatial interaction in urban environments. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 241–248.

[33] Robert Neßelrath, Mohammad Mehdi Moniri, and Michael Feld. 2016. Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions. In *Proceedings of the 12th International Conference on Intelligent Environments*. IEEE, 190–193.

[34] Kai Nickel, Edgar Scemann, and Rainer Stiefelhagen. 2004. 3D-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*. IEEE, 565–570.

[35] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[36] Tony Poitschke, Florian Laquai, Stilyan Stamboliev, and Gerhard Rigoll. 2011. Gaze-based interaction on multiple displays in an automotive environment. In *Proceedings of the International Conference on Systems, Man, and Cybernetics*. IEEE, 543–548.

[37] Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62, 8 (2009), 1457–1506.

[38] David Rempel, Matt J. Camilleri, and David L. Lee. 2014. The design of hand gestures for human-computer interaction: Lessons from sign language interpreters. *International Journal of Human Computer Studies* 72, 10-11 (10 2014), 728–735.

[39] Florian Roider and Tom Gross. 2018. I see your point: Integrating gaze to enhance pointing gesture accuracy while driving. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 351–358.

[40] Florian Roider, Sonja Rümelin, Bastian Pfleging, and Tom Gross. 2017. The effects of situational demands on gaze, speech and gesture input in the vehicle. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 94–102.

[41] Sonja Rümelin, Chadly Marouane, and Andreas Butz. 2013. Free-hand pointing for identification and interaction with distant objects. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 40–47.

[42] Bernhard Scholkopf, Peter L Bartlett, Alex J Smola, and Robert Williamson. 1999. Shrinking the tube: A new support vector regression algorithm. *Advances in Neural Information Processing Systems* (1999), 330–336.

[43] Marco Seeland, Michael Rzanny, Nedal Alaqraa, Jana Wäldchen, and Patrick Mäder. 2017. Plant species classification using flower images - A comparative study of local feature representations. *PLoS ONE* 12, 2 (2 2017), e0170629.

[44] Tevfik Metin Sezgin, Ian Davies, and Peter Robinson. 2009. Multimodal inference for driver-vehicle interaction. In *Proceedings of the 11th International Conference on Multimodal Interfaces*. ACM, 193–198.

[45] Dushyant Kumar Singh. 2015. Recognizing hand gestures for human computer interaction. In *Proceedings of the International Conference on Communications and Signal Processing*. IEEE, 379–382.

[46] Sachin Singh, Niraj Pratap Singh, and Deepti Chaudhary. 2020. A survey on autonomous techniques for music classification based on human emotions recognition. *International Journal of Computing and Digital Systems* 9, 3 (2020), 433–447.

[47] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th ACM International Conference on Multimedia*. ACM, 399–402.

[48] James L. Szalma, Joel S. Warm, Gerald Matthews, William N. Dember, Ernest M. Weiler, Ashley Meier, and F. Thomas Eggemeier. 2004. Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human Factors* 46, 2 (2004), 219–233.

[49] Naohiro Takemura, Toshio Inui, and Takao Fukui. 2018. A neural network model for development of reaching and pointing based on the interaction of forward and inverse transformations. *Developmental Science* 21, 3 (2018).

[50] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2012. Detection of smooth pursuits using eye movement shape features. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 177–180.

[51] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2013. Pursuits: spontaneous interaction with displays based on smooth pursuit eye movement and moving targets. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 439–448.

[52] Qi Ye, Lanqing Yang, and Guangtao Xue. 2018. Hand-free gesture recognition for vehicle infotainment system control. In *Proceedings of the IEEE Vehicular Networking Conference*. IEEE, 1–2.

[53] Yanxia Zhang, Jonas Beskow, and Hedvig Kjellström. 2017. Look but don't stare: Mutual gaze interaction in social robots. In *Social Robotics*. Springer, 556–566.

[54] Dan Zhao, Cong Wang, Yue Liu, and Tong Liu. 2019. Implementation and evaluation of touch and gesture interaction modalities for in-vehicle infotainment systems. In *Image and Graphics*, Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin (Eds.). Springer, 384–394.