Driver Head Pose Estimation with Multimodal Temporal Fusion of Color and Depth Modeling Networks

Susmitha Gogineni and Carlos Busso

***Abstract*— For in-vehicle systems, *head pose estimation* (HPE) is a primitive task for many safety indicators, including driver attention modeling, visual awareness estimation, behavior detection, and gaze detection. The driver's head pose information is also used to augment human-vehicle interfaces for infotainment and navigation. HPE is challenging, especially in the context of driving, due to the sudden variations in illumination, extreme poses, and occlusions. Due to these challenges, driver HPE based only on 2D color data is unreliable. These challenges can be addressed by 3D-depth data to an extent. We observe that features from 2D and 3D data complement each other. The 2D data provides detailed localized features, but is sensitive to illumination variations, whereas 3D data provides topological geometrical features and is robust to lighting conditions. Motivated by these observations, we propose a robust HPE model which fuses data obtained from color and depth cameras (i.e., 2D and 3D). The depth feature representation is obtained with a model based on PointNet++. The color images are processed with the ResNet-50 model. In addition, we add temporal modeling to our framework to exploit the time-continuous nature of head pose trajectories. We implement our proposed model using the *multimodal driving monitoring* (MDM) corpus, which is a naturalistic driving database. We present our model results with a detailed ablation study with unimodal and multimodal implementations, showing improvement in head pose estimation. We compare our results with baseline HPE models using regular cameras, including OpenFace 2.0 and HopeNet. Our fusion model achieves the best performance, obtaining an average *root mean square error* (RMSE) equal to 4.38 degrees.**

I. INTRODUCTION

In the field of *advanced driver-assistance systems* (ADAS), *head pose estimation* (HPE) of the driver is a primitive task for determining several safety metrics for in-vehicle systems. For example, HPE is a key technology for determining driver attention modeling [27], [28]. HPE can also be instrumental for other tasks such as predicting the driver gaze [16], [17], [20], [21], [26] or estimating the drowsiness level of a driver [43]. In addition to solutions to facilitate safety systems, HPE also plays a key role in improving driver-vehicle interfaces for navigation and infotainment purposes [1], [31]. As we transition to autonomous vehicles, it is also important to identify the visual awareness of the driver for take-over tasks [37]. These applications highlight the need for robust in-vehicle solutions for HPE.

Earlier studies in HPE relied on hand-crafted features extracted from 2D color images using classical machine learning methods [33]. With the recent advances in the field of deep learning, neural network-based solutions have been developed to estimate head pose from 2D color images [4], [25], [29], [38], [42]. These methods can be classified into facial landmark-based models [25], [29] and appearance-based models [4], [38], [42]. The latter approach has the advantage of eliminating the dependency on additional pre-processing steps to determine facial landmarks.

The latest developments of *time-of-flight* (ToF) sensors [9] have made it feasible to capture real-time depth data using affordable cameras. The collection of 3D data has promoted the rise of depth-based deep learning methods for HPE. The most commonly used depth formats for HPE are RGB-D [32] and point-cloud [14]. Since head movements are continuous over time, the head pose does not undergo drastic changes from one frame to the next, provided the frames are captured at a sufficiently high frame rate. Thus, examining the continuous trajectory of head pose over time yields valuable insights for assessing the head pose in a new frame. Therefore, temporal modeling is expected to improve HPE. For example, Hu *et al.* [15] demonstrated that *recurrent neural networks* (RNNs) are very effective in capturing the relationship between nearby frames, leading to improvements over a static method that independently processes each frame.

We expect color and depth images to provide unique and complementary features. Figure 1 shows examples of frames collected with color and depth cameras. The 2D color images provide advantages including higher spatial resolution, and details on specific features such as facial landmarks, and face shape. The depth images provide structural and topological details of the face that are not available with color images. In driving scenarios, color images are less robust due to sudden changes in external lighting conditions [18]. Whereas, depth data is more reliable for variable lighting conditions as most of the ToF sensors record distance from time delay of light pulse rather than light intensities. However, depth data is lower in resolution and misses the appearance-based features that can be easily observed in color images. Multimodal computer vision solutions that combine color and depth images can benefit from their advantages, compensating for their individual weaknesses.

This study leverages the complementary features provided by color and depth images, providing a state-of-the-art *multimodal driver head pose estimation* (MD-HPE) approach. The proposed MD-HPE has three main stages. In the first stage, we extract features from the color and point-cloud frames. We use the ResNet-50 architecture [13] to process the color images, and the PointNet++ framework [36] to

The authors are with the Department of Electrical Engineering at the University of Texas at Dallas, Richardson, TX-75080, USA, susmitha.gogineni@utdallas.edu, busso@utdallas.edu

This work was supported by Semiconductor Research Corporation (SRC) Texas Analog Center of Excellence (TxACE) under Grant 2810.014.

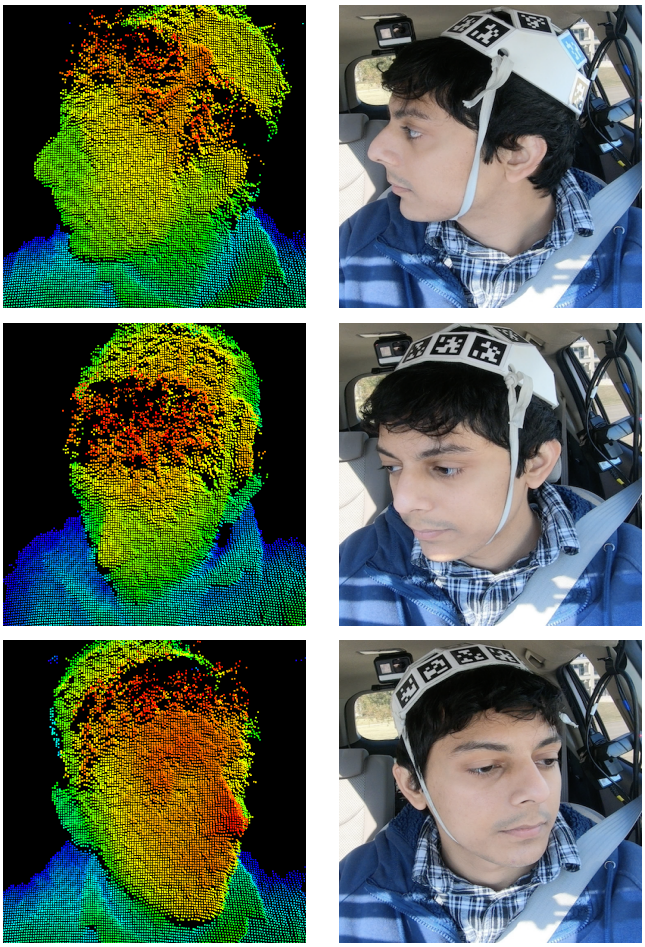

Fig. 1: Examples of frames for one subject in the MDM database [22]. The images on the left are the 3D point-cloud frames. Images on the right are the corresponding 2D color frames. The figure depicts the complementary features provided by the two modalities.

process the point-cloud data for each frame. In the second stage, we model temporal information using *long short-term memory* (LSTM) layers to capture the relationship between head poses across consecutive frames. Then, we concatenate the features from both modalities. In the third stage, we have three individual fully connected layers each corresponding to individual Euler's angles (i.e., yaw, pitch, and roll). The approach is trained by combining the three loss functions for each of the corresponding Euler's angles. The loss function is a combination of classification loss (cross-entropy) and regression loss (mean square error).

We build and test the proposed approach using the *multimodal driver monitoring* (MDM) database [22], which is a naturalistic driving corpus collected from 59 drivers. The ablation study proves that the proposed MD-HPE fusion model performs better than models implemented with individual modalities. We compare the proposed model with baselines relying only on color images, including OpenFace 2.0 [2], HopeNet [38], and a baseline relying only on depth images [15]. The results demonstrate the benefits of combining 2D and 3D images to obtain a robust HPE system. Our best model obtains a *root mean square error* (RMSE) of 4.18° for yaw, 4.29° for pitch, and 4.68° for roll. These results show that our novel approach improves the accuracy, reliability, and robustness of head pose estimation in challenging dynamic settings, such as driving scenarios. The main contributions of our study are:

- We propose to combine features from the color and depth images to robustly estimate head pose using machine-learning strategies. This approach improves accuracy by optimizing the feature set obtained after combining complimenting features from the color and depth modalities.
- We improve the reliability of the HPE algorithm by incorporating the depth modality. This strategy addresses the key challenges of color cameras that are sensitive to varying lighting conditions, sudden movements, and facial occlusions.
- We explore optimal strategies to incorporate temporal modeling in the model, leveraging the relationship between nearby frames.
- We provide extensive evaluations to assess the improvements achieved by the proposed approach, which led to state-of-the-art performance on the MDM corpus.

II. RELATED WORK

A. Head Pose Estimation with Color Images

The computer vision community has extensively studied HPE, where most of these studies are based on color images. Murphy-Chutorian *et al.* [33] and Czuprynski *et al.* [6] provide detailed surveys on studies focusing on HPE. Appearance-based models [38], [42] have recently received increased attention due to the popularity of deep learning. These HPE models offer two advantages. Firstly, they decrease the need for additional pre-processing steps, such as estimating facial landmarks. Secondly, the training set can incorporate synthetic images to enhance accuracy and robustness. Ruiz *et al.* [38] proposed a multi-loss *convolutional neural network* (CNN), which is trained on a large synthetically expanded dataset. Yang *et al.* [42] proposed the FSA-Net, which is based on fine-grained feature aggregation and regression. The FSA-Net framework is a facial landmark free method where the model is trained on spatial relation in the feature map along with features. The feature maps are spatially grouped before aggregation. This grouping is done by defining learnable and non-learnable scoring functions which evaluate the importance of the pixel-level features.

B. Head Pose Estimation with Depth Images

Recent advancements in depth-based sensors have motivated vision researchers to use depth data for perceptual tasks. This strategy is especially important for in-vehicle safety systems where the external visible light conditions are extremely variable, affecting regular color images. Template-matching methods treat the head pose estimation as the registration between the source point-cloud and the reference point-cloud. The head pose value is determined by

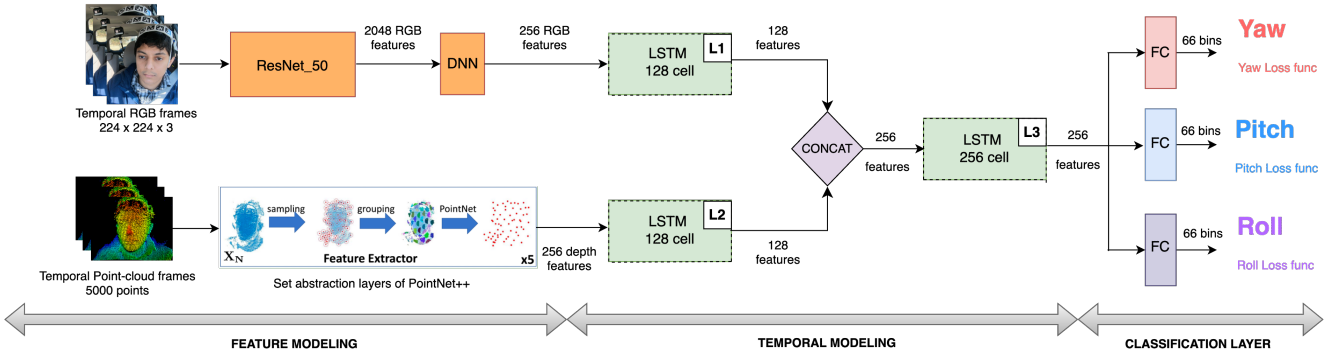

Fig. 2: Proposed multimodal HPE framework. The model includes feature extraction, temporal modeling, and classification layers. The features extracted from the color model (ResNet-50) and point-cloud model (PointNet++) are concatenated. The details of the feature extractor in PointNet++ is shown in Figure 3 and Table I. The temporal modeling is optional and is implemented with the LSTM blocks labeled with L1, L2, and L3 (dashed lines). The model has unique loss values for each of the Euler's angles (yaw, pitch, and roll).

optimizing the registration technique to get the best transformation. For example, Bar *et al.* [3] and Meyer *et al.* [30] used template matching with the *iterative closest point* (ICP) algorithm to estimate head pose from depth data. These template-based methods can be highly sensitive to challenges in driving scenes due to facial occlusions and varying illuminations. These conditions may lead to error in registration and incorrect transformations. Another depth-based HPE approach used regression forests on RGB-D data [8].

Hu *et al.* [14] proposed the first end-to-end deep learning-based HPE method on point-cloud data. This method is based on the PointNet++ framework [36] with a revised set of abstraction layers. PointNet++ has hierarchical abstraction layers to extract both local and global features from the point-cloud data. This approach uses five levels of abstraction layers followed by a 6D regression model. Hu *et al.* [15] further extended their study by implementing a temporal model with *Bidirectional long short-term memory* (BiLSTM) layers after the spatial feature extraction. PointNet++ is a rotation-invariant depth model. State-of-the-art depth models which are rotation equivariant [5], [39] can also be used to solve head pose with respect to a reference frame.

C. Head Pose Estimation by Fusing Color and Depth Images

There are very few methods that explore the fusion of features from multimodal data for HPE. Mukherjee *et al.* [32] estimates head pose from RGB-D data. This method fuses RGB and depth features by taking the average of the class posterior scores from the color and depth classifiers. Li *et al.* [29] fused features from color and sparse point-cloud consisting of 3D-facial landmarks. In this method, features are extracted through *graph convolutional network* (GCN). A limitation of this approach is that the local features from the depth data are ignored and only the global features such as facial landmarks are considered. Wang *et al.* [41] used multi-frame point-cloud registration for HPE. In this method, the gaze region is also estimated by using the segmented eye region of the color image and the estimated head pose. Our proposed MD-HPE approach is different from these studies, since (1) it strategically combines at the model level information from color and depth cameras, using both information to predict head pose, (2) it leverages the entire information conveyed in the color image and the point-cloud data, (3) it captures temporal information to exploit the relation between nearby frames, and (4) it demonstrates its performance on a naturalistic driving database achieving state-of-the-art HPE performance.

III. PROPOSED APPROACH

This paper proposes a fusion network for HPE, which extracts complimentary features from the color images and point-cloud data. Figure 2 shows the overview of the proposed network. The network contains three main stages including feature extraction, temporal modeling, and classification layers. This section presents details about the fusion network and the loss function.

A. Feature Modeling

In the feature extraction stage, the color frames are modeled with the ResNet-50 network [13], which is pre-trained with the ImageNet database [7]. The output of the ResNet-50 network is a 2,048-feature vector, which serves as one of the inputs of our model. The next block is a *fully-connected* (FC) layer with 256 nodes that project the feature vector into a reduced embedding.

For point-cloud data, we rely on the PointNet++ framework [36] to obtain a discriminative feature representation from the depth data. PointNet++ extracts both fine-grained local and global features from depth data. In PointNet++, the feature extraction layer is referred to as *set abstraction layer*. In our network, we implement five set abstraction layers in series. Initial layers extract local features and the later layers extract global features from the point-cloud. Each set abstraction layer is constructed with three blocks: sampling, grouping, and PointNet as shown in Figure 3.

TABLE I: Implementation details for the five set abstraction layers (sampling, grouping and PointNet layers). m represents the number of anchor points that are chosen for the sampling operation. $d_c = 3$ is the dimension of the input coordinate system (i.e., x, y, and z). r represents the radius of the ball query for the grouping operation. l represents the number of points within r to be chosen in the grouping step. d_f is the feature dimension of the previous set. For the first set, d_f=d_c=3.

Layer	Specification	Output Dimension
Sampling	m=512	[m=512,d_c=3] (**S1**)
Grouping	[r=0.1,l=4]	[m=512,d_f=3,l=4] (**G1**)
	[r=0.2,l=8]	[m=512,d_f=3,l=8] (**G2**)
	[r=0.4,l=16]	[m=512,d_f=3,l=16] (**G3**)
PointNet	[8,8,16]	[512,16] (**T1**)
	[16,16,32]	[512,32] (**T2**)
	[16,24,32]	[512,32] (**T3**)
		[512,80] (**T4**)
Sampling	m=256	[m=256,d_c=3]
Grouping	[r=0.15,l=8]	[m=256,d_f=83,l=8] (**G4**)
	[r=0.3,l=16]	[m=256,d_f=83,l=16] (**G5**)
	[r=0.5,l=24]	[m=256,d_f=83,l=24] (**G6**)
PointNet	[16,16,64]	[256,64]
	[32,32,64]	[256,64]
	[48,48,96]	[256,96]
		[256,224]
Sampling	m=128	[m=128,d_c=3]
Grouping	[r=0.15,l=8]	[m=128,d_f=227,l=8]
	[r=0.3,l=16]	[m=128,d_f=227,l=16]
	[r=0.5,l=24]	[m=128,d_f=227,l=24]
PointNet	[16,16,64]	[128,64]
	[32,32,64]	[128,64]
	[48,48,96]	[128,96]
		[128,224]
Sampling	m=64	[m=64,d_c=3]
Grouping	[r=0.2,l=16]	[m=64,d_f=227,l=16]
	[r=0.4,l=32]	[m=64,d_f=227,l=32]
	[r=0.8,l=48]	[m=64,d_f=227,l=48]
PointNet	[32,32,64]	[64,64]
	[48,48,64]	[64,64]
	[64,64,128]	[64,128]
		[64,256]
Sampling	m=1	[m=1,d_c=3]
Grouping	[r=inf]	[64,259]
PointNet	[256,256,256]	[256] (**T5**)

For the sampling block, the method selects a predefined number (m) of anchor points from the point-cloud data using the *iterative farthest point sampling* (IFPS) algorithm. The set with the sampled anchor point is denoted by S. For the grouping block, the goal is to group neighboring points to the anchors. The approach relies on the ball query algorithm. Since the density of points in the point-cloud is not uniformly distributed, the approach extracts robust features using the *multi-scale grouping* (MSG) method [36]. In the MSG method, grouping is done at multiple scales of resolution. Our implementation uses three different scales of resolutions: G1, G2, and G3 on each of the anchor points (see Table I for the parameters used in our model). For the PointNet block, the goal is to obtain local patterns that are discriminative for the HPE task. PointNet captures point-to-point relations in a local region and transforms features into higher-dimension representations. The PointNet [35] consists of three *multilayer perceptron* (MLP) layers with shared

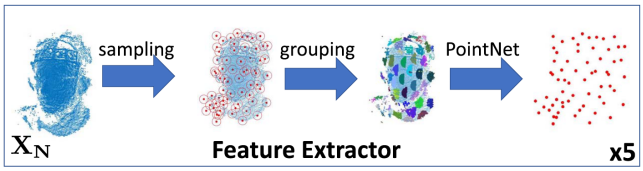

Fig. 3: Set abstraction layer in the PointNet++ [36], including sampling, grouping, and PointNet layers. The framework directly uses point-cloud data without projecting the 3D points into 2D spaces.

weights and a max pooling at the end. At each anchor point, the three multi-scale groups are transformed by the PointNet network into higher dimension features (T1, T2, T3). These multi-scale group features (T1, T2, T3) are concatenated into a single feature set (T4), as implied in Table I. In our model, we also use a series of five set abstraction layers. Similar to CNN, the initial layers extract local features and later layers add global features. In the series of five set abstraction layers, the output from the PointNet layer of an abstraction layer will be the input for the sampling layer in the next abstraction layer. The first four abstraction layers follow the same implementation with different sampling sizes and grouping scales (see Table I). However, the last abstraction layer has only one anchor point in the sampling layer. In the grouping layer, all points are grouped together with an infinite radius. In the PointNet layer, weighted average pooling is used instead of max pooling. The final feature set (T5), is a 256-D vector. Table I shows the implementation details of each set abstraction layer.

B. Temporal Modeling

In our approach, we employ *long short-term memory* (LSTM) layers to capture temporal information from consecutive frames. However, we need to experiment with the optimal placement of these LSTM layers in our implementation since it remains unclear the optimal placement. Figure 2 shows our model, which includes three LSTM blocks referred to as L1, L2, and L3. After the feature extraction layers, the L1 and L2 blocks process the color and depth modalities, respectively. The L1 and L2 blocks each consists of 128 neurons. The L3 block is implemented after concatenating the feature representations from both modalities, and it encompasses 256 neurons to account for the concatenated feature representations. These three LSTM blocks are denoted by dashed lines to indicate their optional usage. In Section V, we will evaluate various configurations to determine the most effective placement of these LSTM layers. In scenarios where L1 and L2 are not employed (i.e., when there is no temporal model or only L3 is utilized), we concatenate the 256-dimensional color and 256-dimensional depth features to create a 512-dimensional feature vector.

C. Classification Layers

After the LSTM, the feature representation is used as input to three separate FC layers to predict the Euler's angles (i.e., yaw, pitch, and roll). Inspired by the multi-loss function

used in the fine-grained HPE approach proposed by Ruiz *et al.* [38], we solve this problem with a combination of classification and regression losses. The classification loss is formulated with the *cross entropy* (CE) loss. The estimated head pose is assumed to be in the range from -99° to +99°. We split the angles into 66 bins, where each bin corresponds to an angle of 3°. The FC layer does multi-class classification with a softmax activation. The output of the FC layer is the probability distribution across the 66 classification bins. The cross-entropy is calculated between the probabilistic distributions of labeled and predicted bins. The regression loss is computed with the *mean square error* (MSE) loss between the labeled and predicted angles. The final loss function is formulated as a linear combination of the classification loss ($\mathcal{L}_{CE}$) and regression loss ($\mathcal{L}_{MSE}$), using the scaling factor α:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{MSE} \tag{1}$$

The idea behind the multi-loss approach is to guide the model to predict the neighboring angles for the head pose using the classification loss. Then, the regression loss leads the model to predict the fine-grained pose. The error is independently back-propagated for each branch since we use three separate FC layers for yaw, pitch, and roll.

IV. EXPERIMENTAL SETTINGS

This section discusses the experimental settings, starting with the details of the MDM database, which is used to train and assess the proposed approach. We also describe the ground truth labeling and calibration approaches required for the evaluation. Additionally, we provide the details on the infrastructure used to conduct the evaluation.

A. MDM Database

We train the network with the *multimodal data monitoring* (MDM) database [22], which is a naturalistic driving corpus collected from 59 gender-balanced subjects. The data includes a set of in-vehicle sensors, including four RGB cameras, a *time-of-flight* (ToF) camera, and a microphone array. Additionally, a *controller area network bus* (CAN-Bus) is used to record the dynamic information of the vehicle. The 2D color data is collected with four GoPro HERO6 cameras placed to record (1) the frontal view of the driver's face, (2) the semi-profile view of the driver's face as viewed from the rear mirror, (3) the back side of the driver's head to record the Fi-Cap helmet [19], and (4) the road view. The color data is collected at a frame rate of 60 *frame per second* (FPS) at a resolution of 1920×1080. We implement our color-based model with the data obtained with the frontal view camera. The depth data is collected with a PMD Picoflexx camera, which recorded the frontal view of the driver's face. Picoflexx captures both point-cloud and gray-scale infrared frames at a maximum frame rate of 45 FPS and a resolution of 224×171. Figure 1 shows an example of the color and point-cloud frames for a subject. More details on the data collection protocol are given in Jha *et al.* [22].

The key feature of the corpus is that the ground truth head-pose labels can be determined for continuous frames by tracking the Fi-Cap helmet [19]. The Fi-Cap is a helmet with 23 predefined 2D fiducial markers called AprilTags [34]. This helmet can be partially seen in Figure 1, and it is exclusively used to extract the ground truth (i.e., it is not used during inference). The Fi-Cap helmet is worn on the back of the driver's head, preventing face occlusions when recorded with a frontal view. The camera behind the driver recorded the Fi-Cap helmet for all the frames. The ground truth head-pose labels are determined by comparing the orientation and position of Fi-Cap helmet between the selected and reference frames. The position of these 2D markers can be detected with vision algorithms [40]. The Kabsch algorithm [23] is used to determine the relative orientation and position of the Fi-Cap helmet in each intermediate frame with respect to a reference frame. Jha and Busso [19] provide more details on how to extract the head-pose of the driver from the Fi-Cap helmet.

Each of the subjects might wear the Fi-Cap helmet in relatively different orientations. Moreover, within a subject's recording, the Fi-Cap orientation can slightly change during the recordings due to the continuous movement of the vehicle and the driver's head. Therefore, we need a calibration process for the ground truth labels to be defined uniformly across and within the subjects. We use the multiple *local reference frame* (LRF) calibration approach proposed by Hu *et al.* [15]. In this calibration process, one of the frontal frames from one subject is manually selected to be the *global reference* (GR) frame. The head pose of the GR frame is estimated using OpenFace 2.0 [2]. For each subject, a number of frames that have the closest rotation angles to the GR frame are chosen as the *local reference* (LR) frames. The rotation matrix of the LR is denoted as R_{LRF}. For each LR frame, the face region in the point-cloud data for the LR and GR frames is cropped. Then, the ICP algorithm is run between the GR and LR point-cloud data to find a transformation between them. This transformation matrix ($R_{LocalGlobal}$) is used to compensate for the differences in Fi-Cap orientation between the LR and GR frames. Finally, the calibration matrix (R_c) for each LR is determined as $R_c = R_{LocalGlobal}R_{LRF}$. For a frame at time t, with rotation R_{FiCap}, the closest LR is determined by the timestamp, and the final head-pose rotation is calibrated as $R_t = R_c^{-1}R_{FiCap}$.

$$R_c = R_{LocalGlobal}R_{LRF} \tag{2}$$

$$R_t = R_c^{-1}R_{FiCap} \tag{3}$$

As the color and depth sensors record frames at different frame rates, a clapping board was used at the beginning of the recording to temporally synchronize different sensors. We align the intermediate frames based on the time elapsed from the reference clapping frame to the current frame. In the pre-processing of the color frames, we crop the original image leaving only the face region with a 2D bounding box determined by the OpenFace 2.0 toolkit [2]. The smallest

TABLE II: Ablation study without temporal modeling (e.g., without LSTM layers). This table compares the performance of the HPE methods implemented with (1) only color images, (2) only depth data, and (3) color and depth information.

Modality	Error	Y (°)	P (°)	R (°)	Mean (°)
RGB	RMSE	5.48	5.72	5.98	5.72
	MAE	3.85	4.18	4.54	4.19
Depth	RMSE	6.44	5.50	5.74	5.89
	MAE	4.14	3.75	4.05	3.98
Fusion	RMSE	**5.08**	**4.65**	**4.92**	**4.88**
	MAE	**3.73**	**3.32**	**3.77**	**3.60**

side of the cropped image is resized to 224 and the biggest side is resized to 224 multiplied by the aspect ratio of the bounding box. Next, a random crop is done such that the final image size is 224×224. With this approach, the image is resized to 224×224 while preserving the aspect ratio of the bounding box for the face. Finally, the image intensities are normalized per channel.

For point-cloud data, we use distance-based and statistical filters to remove points that are clearly part of the background. Then, we use a voxel grid downsampling approach to sample 5,000 points from each of the point-cloud frames. Each point-cloud frame is normalized such that the centroid is at the origin, and all the points lie inside a unit sphere. The dataset is partitioned into a train set (39 drivers), a development set (10 drivers), and a test set (10 drivers).

B. Model Settings

The model is trained using the ADAM optimizer [24] with initial learning rate set to 0.001 and a learning rate decay of 0.7 per 75,000 steps. The scaling factor α is set to 0.1 in the loss function (Eq. 1). The fusion model is implemented using TensorFlow and trained using an NVIDIA GeForce RTX 3090 Ti GPU.

V. EXPERIMENTAL RESULTS

In this section, we provide a detailed ablation study of our model, and we compare with state-of-the-art baselines. In our study, we use the *root mean square error* (RMSE) and the *mean absolute error* (MAE) as the metrics to evaluate the performance.

A. Ablation Study

We perform extensive experiments to analyze the performance improvement attained by our model when we (1) fuse depth and color images, and (2) add temporal modeling layers. Table II shows the HPE results when using color, depth, and the combination of both modalities. For this analysis, we do not include any LSTM blocks, independently estimating the results for each frame. The table shows that the errors consistently drop when jointly using both modalities. This result shows that the valuable complementary features offered by incorporating both color and depth modalities can improve HPE.

Table III shows the results when adding the LSTM blocks. We evaluate different configurations. The first part of the table reports the results obtained by adding either L1 or L2 to the unimodal systems. When we compare the results without a temporal model (Table II) and with a temporal model (Table III), we observe improvements, indicating that leveraging the relationship between consecutive frames is important. Table III also compares the HPE performance achieved when adding L1 and L2, and when adding only L3. The most effective strategy is adding the LSTM block after the concatenation (i.e., L3). This model outperforms all other implementations of our framework.

Figure 4 shows six examples with actual and predicted head poses using the proposed model.

B. Comparison with Baselines

We compare our proposed model to several unimodal baselines. We use our model implemented with the L3 block.

TABLE III: Ablation study comparing the performance of the HPE method with temporal modeling implemented with (1) only color images followed by the L1 LSTM layer, (2) only depth data followed by the L2 LSTM layer, (3) fusion without temporal modeling, (4) fusion followed by the L1 and L2 LSTM layers, and (5) fusion followed by the L3 LSTM layer.

Modality	L1	L2	L3	Error	Y (°)	P (°)	R (°)	Mean (°)
RGB	✓	✗	✗	RMSE	4.38	6.10	6.52	5.60
				MAE	2.92	4.52	4.91	4.11
Depth	✗	✓	✗	RMSE	4.94	5.04	5.21	5.06
				MAE	3.54	3.85	3.56	3.65
Fusion	✗	✗	✗	RMSE	5.08	4.65	4.92	4.88
				MAE	3.73	3.32	3.77	3.60
Fusion	✓	✓	✗	RMSE	4.57	5.36	5.56	5.16
				MAE	3.32	4.18	4.33	3.94
Fusion	✗	✗	✓	RMSE	**4.18**	**4.29**	**4.68**	**4.38**
				MAE	**3.02**	**3.32**	**3.42**	**3.25**

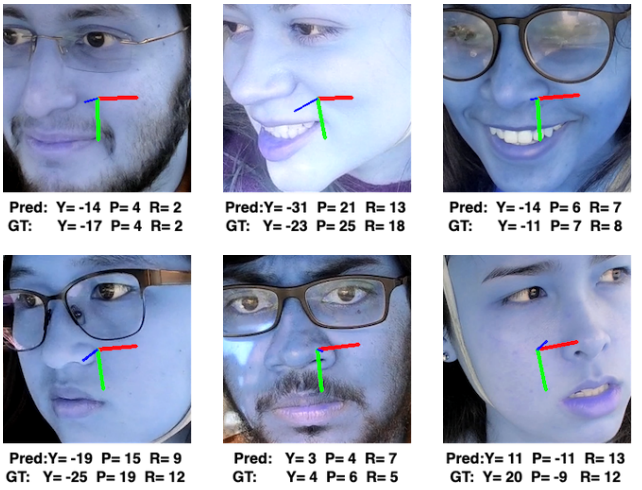

Fig. 4: Examples of HPE results obtained with the proposed multimodal method displayed on the 2D images. The blue axis points towards the front of the face, the green axis points downward and the red axis points to the side (*Pred*: predicted angles, *GT*: ground truth angles).

TABLE IV: Comparison of proposed HPE model with unimodal baselines. The table shows results using the RMSE metric.

Model	Y (°)	P (°)	R(°)	Mean(°)
OpenFace 2.0 [2]	5.06	7.20	7.63	6.63
HopeNet [38]	6.29	8.09	6.62	7.0
Hu *et al.* [15]	4.69	5.30	5.63	5.20
Proposed Model	**4.18**	**4.29**	**4.68**	**4.38**

For this analysis, we only consider RMSE results. First, we consider color-based HPE using OpenFace 2.0 [2], and HopeNet [38]. Table IV shows that the proposed multimodal model provides better predictions than these systems. Second, we compare our approach with the temporal model presented by Hu *et al.* [15], which uses only point-cloud data. This model is better than the color-based baselines. However, our multimodal approach clearly leads to better results. The key advantage of our approach is the use of color and point-cloud images, leveraging the complementary features provided by 2D and 3D images.

VI. CONCLUSIONS

This paper proposed an HPE model that effectively combines color images and depth data. The multimodal approach takes advantage of the spatial resolution provided by color images and the structural information provided by the depth data. The approach also incorporates temporal modeling leveraging the relationship between the head pose across nearby frames. It demonstrates high accuracy and robustness in naturalistic in-vehicle recordings. Experimental evaluation indicates that the fusion model outperforms similar approaches implemented solely with color or depth information. Furthermore, it achieves improvements over state-of-the-art unimodal baselines for HPE.

As part of our future work, we are planning to improve the model to achieve robustness even when some modalities are missing. Studies in other multimodal problems have demonstrated that this is possible [10]–[12]. This is important since in-vehicle scenarios often include cases with missing or corrupted features (e.g., saturated color images due to extreme illumination, occlusion of the face due to steering wheel operation). In such scenarios, we still need the model to be robust and reliable. Another promising future direction is to enhance this framework by using rotation equivariant models to process the color and depth images.

REFERENCES

[1] A. Aftab, M. von der Beeck, and M. Feld. You have a point there: Object selection inside an automobile using gaze, head pose and finger pointing. In *ACM International Conference on Multimodal Interaction (ICMI 2020)*, pages 595–603, Virtual Event, October 2020.

[2] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 59–66, Xi'an, China, May 2018.

[3] T. Bär, J. Reuter, and J. Zöllner. Driver head pose and gaze estimation based on multi-template ICP 3-D point cloud alignment. In *International IEEE Conference on Intelligent Transportation Systems (ITSC 2012)*, pages 1797–1802, Anchorage, AK, USA, September 2012.

[4] F. J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. FacePoseNet: Making a case for landmark-free face alignment. In *IEEE International Conference on Computer Vision Workshops (ICCVW 2017)*, pages 1599–1608, Venice, Italy, October 2017.

[5] H. Chen, S. Liu, W. Chen, H. Li, and R. Hill. Equivariant point network for 3D point cloud analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, pages 14509–14518, Nashville, TN, USA, June 2021.

[6] B. Czupryński and A. Strupczewski. High accuracy head pose tracking survey. In D. Ślezak, G. Schaefer, S. Vuong, and Y. Kim, editors, *International Conference on Active Media Technology (AMT 2014)*, volume 8610 of *Lecture Notes in Computer Science*, pages 407–420. Springer Berlin Heidelberg, Warsaw, Poland, August 2014.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, USA, June 2009.

[8] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 617–624, Providence, RI, USA, June 2011.

[9] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (ToF) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, September 2011.

[10] L. Goncalves and C. Busso. AuxFormer: Robust approach to audiovisual emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 7357–7361, Singapore, May 2022.

[11] L. Goncalves and C. Busso. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170, October-December 2022.

[12] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso. Versatile audiovisual learning for handling single and multi modalities in emotion regression and classification tasks. *ArXiv e-prints (arXiv:2305.07216)*, pages 1–14, May 2023.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778, Las Vegas, NV, USA, June-July 2016.

[14] T. Hu, S. Jha, and C. Busso. Robust driver head pose estimation in naturalistic conditions from point-cloud data. In *IEEE Intelligent Vehicles Symposium (IV 2020)*, pages 1176–1182, Las Vegas, NV USA, October-November 2020.

[15] T. Hu, S. Jha, and C. Busso. Temporal head pose estimation from point cloud in naturalistic driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8063–8076, July 2022.

[16] S. Jha, N. Al-Dhahir, and C. Busso. Driver visual attention estimation using head pose and eye appearance information. *IEEE Open Journal of Intelligent Transportation System*, 4:216–231, March 2023.

[17] S. Jha and C. Busso. Analyzing the relationship between head pose and gaze to model driver visual attention. In *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, pages 2157–2162, Rio de Janeiro, Brazil, November 2016.

[18] S. Jha and C. Busso. Challenges in head pose estimation of drivers in naturalistic recordings using existing tools. In *IEEE International Conference on Intelligent Transportation (ITSC 2017)*, pages 1624–1629, Yokohama, Japan, October 2017.

[19] S. Jha and C. Busso. Fi-Cap: Robust framework to benchmark head pose estimation in challenging environments. In *IEEE International Conference on Multimedia and Expo (ICME 2018)*, pages 1–6, San Diego, CA, USA, July 2018.

[20] S. Jha and C. Busso. Probabilistic estimation of the gaze region of the driver using dense classification. In *IEEE International Conference on Intelligent Transportation (ITSC 2018)*, pages 697–702, Maui, HI, USA, November 2018.

[21] S. Jha and C. Busso. Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions. *IEEE Transactions on Intelligent Vehicles*, 8(1):59–72, January 2023.

[22] S. Jha, M. Marzban, T. Hu, M. Mahmoud, N. Al-Dhahir, and C. Busso. The multimodal driver monitoring database: A naturalistic corpus to study driver attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10736–10752, August 2022.

[23] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A.*, A34(Part 5):827–828, September 1978.

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13, San Diego, CA, USA, May 2015.

[25] A. Kumar, A. Alavi, and R. Chellappa. KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pages 258–265, Washington, DC, USA, May-June 2017.

[26] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim. Real-time gaze estimator based on driver's head orientation for forward collision warning system. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):254–267, March 2011.

[27] N. Li and C. Busso. Analysis of facial features of drivers under cognitive and visual distractions. In *IEEE International Conference on Multimedia and Expo (ICME 2013)*, pages 1–6, San Jose, CA, USA, July 2013.

[28] N. Li, J. Jain, and C. Busso. Modeling of driver behavior in real world scenarios using multiple noninvasive sensors. *IEEE Transactions on Multimedia*, 15(5):1213–1225, August 2013.

[29] Y. K. Li, Y. Z. Yu, Y. L. Liu, and C. Gou. MS-GCN: Multi-stream graph convolution network for driver head pose estimation. In *IEEE International Conference on Intelligent Transportation Systems (ITSC 2022)*, pages 3819–3824, Macau, China, October 2022.

[30] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz. Robust model-based 3D head pose estimation. In *IEEE International Conference on Computer Vision (ICCV 2015)*, pages 3649–3657, Santiago, Chile, December 2015.

[31] T. Misu. Visual saliency and crowdsourcing-based priors for an in-car situated dialog system. In *International conference on Multimodal interaction (ICMI 2015)*, pages 75–82, Seattle, WA, USA, November 2015.

[32] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, November 2015.

[33] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, April 2009.

[34] E. Olson. AprilTag: A robust and flexible visual fiducial system. In *IEEE International Conference on Robotics and Automation (ICRA 2011)*, pages 3400–3407, Shanghai, China, May 2011.

[35] C. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 77–85, Honolulu, HI, USA, July 2017.

[36] C. Qi, L. Yi, H. Su, and L. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *In Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5099–5108, Long Beach, CA, USA, December 2017.

[37] Y. Qiu, C. Busso, T. Misu, and K. Akash. Incorporating gaze behavior using joint embedding with scene context for driver takeover detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 4633–4637, Singapore, May 2022.

[38] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2018)*, pages 2187–2196, Salt Lake City, UT, USA, June 2018.

[39] H. Wang, Y. Liu, Z. Dong, and W. Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *ACM International Conference on Multimedia (MM 2022)*, pages 1630–1641, Lisboa, Portugal, October 2022.

[40] J. Wang and E. Olson. AprilTag 2: Efficient and robust fiducial detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016)*, pages 4193–4198, Daejeon, South Korea, October 2016.

[41] Y. Wang, G. Yuan, and X. Fu. Driver's head pose and gaze zone estimation based on multi-zone templates registration and multi-frame point cloud fusion. *Sensors*, 22(9):3154, April 2022.

[42] T. Y. Yang, Y. T. Chen, Y. Y. Lin, and Y. Y. Chuang. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 1087–1096, Long Beach, CA, USA, June 2019.

[43] M. Ye, W. Zhang, P. Cao, and K. Liu. Driver fatigue detection based on residual channel attention network and head pose estimation. *Applied Sciences*, 11(19):9195, October 2021.