2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Probabilistic Estimation of the Gaze Region of the Driver using Dense Classification*

Sumit Jha and Carlos Busso

Abstract— **The ability to monitor the visual attention of a driver is a useful feature for smart vehicles to understand the driver's intents and behaviors. The gaze angle of the driver is not deterministically related to his/her head pose due to the interplay between head and eye movements. Therefore, this study aims to establish a probabilistic relationship using deep learning. While probabilistic regression techniques such as *Gaussian process regression* (GPR) has been previously used to predict the visual attention of a driver, the proposed deep learning framework is a more generic approach that does not make assumptions, learning the relationship between gaze and head pose from the data. In our formulation, the continuous gaze angles are converted into intervals and the grid of the quantized angles is treated as an image for dense prediction. We rely on *convolutional neural networks* (CNNs) with upsampling to map the six degrees of freedom of the orientation and position of the head into gaze angles. We train and evaluate the proposed network with data collected from drivers who were asked to look at predetermined locations inside a car during naturalistic driving recordings. The proposed model obtains very promising results, where the size of the gaze region with 95% accuracy is only 11.73% of a half sphere centered at the driver, which approximates his/her field of view. The architecture offers an appealing and general solution to convert regression problems into dense classification problems.**

I. Introduction

The visual attention of a driver is an important factor to understand his/her mental state and his/her ability to perform relevant actions related to the driving task. Drivers obtain most of the information to operate a vehicle through vision and their inability to spot a potentially dangerous situation such as a pedestrian or another vehicle on the road can lead to unfortunate accidents. The ability of a smart vehicle to track where the driver is looking can be a useful feature for *advanced driver-assistance systems* (ADAS). In an autonomous driving scenario, the knowledge of the driver's visual attention can help a smart vehicle to take decisions on behalf of a distracted driver or hand over control to an attentive driver when the machine is unable to make a complex maneuvering decision.

An accurate system to track the driver's gaze requires detailed information about the eyes and their pupils. Detecting gaze requires very specific sets of sensors to efficiently perform this task [1], [2]. However, the car environment brings important challenges to complete this task due to changes in illumination, occlusions and extreme head rotations. An alternative approach is to approximate gaze direction with the head pose which is easier to track [3], [4]. The knowledge of the head pose of a driver can be helpful in providing useful information about the driver's visual attention [5]. While a deterministic relationship does not exist between the head pose and the gaze, there is a strong correlation which can be leveraged to infer important information about the driver's visual attention. We have argued that detecting confidence regions containing the gaze of the driver using a probabilistic framework can be more effective than predicting an intrinsic noisy single gaze point [6].

Jha and Busso [6] proposed a probabilistic method to predict a confidence region for the gaze of a driver given his/her head pose. The framework relied on the *Gaussian process regression* (GPR) algorithm to predict a Gaussian distribution of gaze conditioned on the head pose. The GPR framework assumes that the outputs are samples generated from a Gaussian process that is dependent on the input. The output of this framework is limited by the Gaussian assumptions. This study explores a more flexible framework that does not make any assumption, learning the relation between head pose and gaze from the data. Recent studies have suggested that using classification on discretized intervals provides a more generic approach than predicting a probabilistic map [7]. This paper leverages this idea with deep learning models. The advances in deep learning provide an appealing approach to complete this task with dense neural networks. Using the head pose of the driver as input, we formulate this task as a classification problem by discretizing the range of gaze angles. The grid of various possible horizontal and vertical gaze angles are treated as an image and a *convolutional neural network* (CNN) is used to infer the probabilities at different gaze angles. The resolution of the grid is gradually increased so that the resolution and precision of the estimation increase as the network gets deeper.

The experimental evaluation provides strong evidences of the capabilities of the proposed model. We can obtain saliency regions of arbitrary shapes that are learned from the data. The results provide clear evidences of the benefits of the proposed architecture where the size of the gaze region with 95% accuracy is only 11.73% of a half sphere centered at the driver, which approximates his/her field of view. This model provides an appealing solution to formulate regression problems as classification tasks using methods for dense prediction.

II. Related Work

Estimating the visual attention of a driver from his/her head pose is an important problem. Several studies have proposed different solutions [6], [8], [9]. Head pose only

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

The authors are with the Multimodal Signal Processing Lab at the University of Texas at Dallas, Richardson, TX-75080, USA sumit.jha@utdallas.edu, busso@utdallas.edu

978-1-7281-0322-8/18/$31.00 ©2018 IEEE

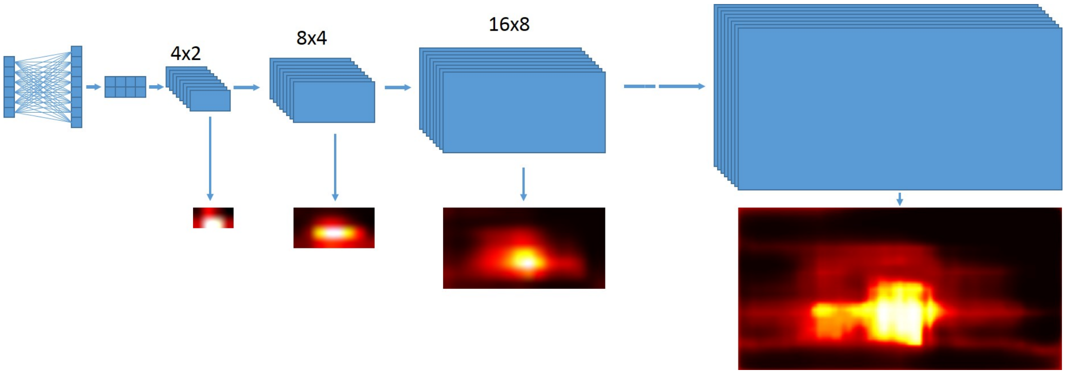


Fig. 1. The proposed architecture. An image is generated after each upsampling stage. The output at each stage is optimized through back propagation. There are seven stages: 4×2 (before upsampling), 8×4, 16×8, 32×16, 64×32, 128×64 and 256×128. The figure shows examples of the predicted gaze regions.

provides partial information about the gaze, so accurate gaze cannot be predicted solely from the head pose. When a driver glances at an object, his/her head and eye move to reach the target point. The interaction between head and eye movements make the mapping between head pose and gaze non-deterministic [5]. Therefore, many researchers simplify the problem by predicting partial gaze information such as gaze zones [9], [10], studying eyes-off-the-road events [11], [12] or detecting primary driving actions such as mirror-checking actions [13]. Jha and Busso [6] proposed an alternative approach where they predict a confidence region of gaze depending on the head pose. The approach was a Gaussian model for the gaze conditioned on the head pose. Each head pose predicted a Gaussian distribution of possible gaze directions. The distribution was used to define the confidence region for the gaze region. This paper proposes a more generic approach based on deep learning that can predict an arbitrary distribution without compromising the accuracy and resolution of the models.

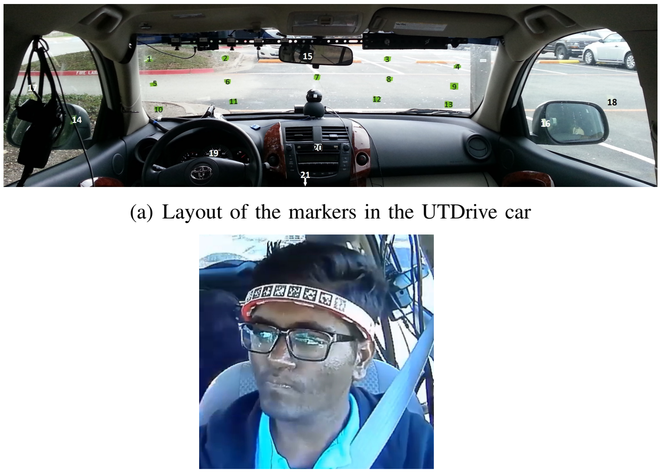
There has been increasing interest in treating a regression task as a classification problem by discretizing continuous labels. Recent studies related to image and sound generation [7], [14] have suggested that a better approach to create an arbitrary probability distribution is by using a classifier with softmax activation. This approach is better than regression techniques such as *mixture density network* (MDN) [15]. We argue that this approach can provide a better way to learn an arbitrary, non-parametric gaze distribution from head pose. Torgo and Gama [16] discussed different methods to divide the regression scores into intervals based on priori information of the output and the underlying domain. They used an iterative approach to obtain the most optimum intervals for a given classification algorithm to get the best regression accuracy. Frank and Hall [17] suggested an approach to implicitly consider ordered classification problems, where the labels represent intervals (such as cold, warm and hot). They used decision trees to hierarchically split the boundaries into different regions. Le et al. [18] suggested the use of k-means clustering to discretize continuous labels using the dichotomized labels to predict continuous emotional labels (valence, arousal and dominance). Since the classes need to be ordered, they use cost-sensitive cross entropy as the loss function.

Dense prediction has been studied mostly in the field of image generation, image segmentation [19], and autoencoders [20], where each pixel is treated as an output node. In most of these applications, the inputs and outputs of the network are both images, so the architecture has either pooling steps followed by unpooling operations [19], [21], or a receptive field which is expanded using dilation [22]. A few studies have considered inverting a CNN for autoencoders and semantic segmentations [19]–[21]. Since the pooling layer is non-invertible, Zeiler et al. [21] suggested an unpooling layer that approximates the inverse of a pooling operation. It stores the pixel location when pooling so that they can restore the value to this location when unpooling. All other pixels are assigned to zero. This approach was also used by Noh et al. [23] to perform semantic segmentation, and by Zeiler et al. [20] to regenerate an image from a CNN to visualize and understand the feature maps. Long et al. [19] used upsampling to generate a semantic segmentation map. Since the interpolation function is learned using convolution layers, they can learn any non-linear interpolation.

These deep learning studies offer appealing solutions to our problem, where we aim to generate a visual map to characterize the visual attention of the driver from his/her head pose. To the best of our knowledge, this is the first study to explore this idea, opening novel research direction in this area.

III. Method

Our aim is to quantize the gaze angle values into fixed intervals such that the regression problem can be reformulated as a multi-class classification problem. We transform the possible horizontal and vertical gaze angles into grids of equal width, representing the gaze location as an image. We use an inverted CNN architecture to gradually upsample the image to increase the resolution of the predicted salient map region that describes the visual attention of the driver. Figure 1 shows the proposed architecture, which is explained in this section. We use a technique inspired by transfer learning [24] and forward thinking [25] to train our model. We use upsampling layers which repeats the pixel values from a



(a) Layout of the markers in the UTDrive car

(b) Driver wearing headband

Fig. 2. (a) Markers on the UTDrive car, (b) a driver with the headband.

TABLE I

PROPOSED DEEP LEARNING ARCHITECTURE.

Layer	Spec	Activation	Dropout	Output Dimension
Dense	**8**	**ReLU**	**0.5**	**1 x 8**
Reshape	**4 x 2**	**-**	**-**	**1 x 4 x 2**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 4 x 2**
Upsampling2D	**2x2**	**-**	**-**	**8 x 8 x 4**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 8 x 4**
Upsampling2D	**2x2**	**-**	**-**	**8 x 16 x 8**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 16 x 8**
Upsampling2D	**2x2**	**-**	**-**	**8 x 32 x 16**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 32 x 16**
Upsampling2D	**2x2**	**-**	**-**	**8 x 64 x 32**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 64 x 32**
Upsampling2D	**2x2**	**-**	**-**	**8 x 128 x 64**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 128 x 64**
Upsampling2D	**2x2**	**-**	**-**	**8 x 256 x 128**
Conv2D	**8, 3x3**	**ReLU**	**0.5**	**8 x 256 x 128**
Conv2D	**1, 3x3**	**Softmax**	**0.5**	**1 x 256 x 128**

previous layer, followed by convolution layers to perform the interpolation operation. The system is retrained after adding a new upsampling layer, where the entire network is tuned using back propagation.

As we increase the resolution of the output, the number of classes becomes unmanageable. For example, if we maintain the range of horizontal angles between -110° and +110° and the range of vertical angles between -100° and +10°, we would need 24,200 classes to obtain a 1° resolution. Furthermore, we also need to maintain spatial consistency in the output, where a confusion between (1,1) and (100,100) should be penalized more than a confusion between (99,99) and (100,100). We address these problems by using CNNs with gradual upsampling.

A. *Network Architecture*

This section explains the overall architecture that takes the head pose characterized by a 6D input vector (three head position coordinates and three head rotation angles) and provides an output map of 256x128 pixels. Figure 1 describes the architecture, which is inspired by a deconvolution network that inverts a CNN [19]. We start with a fully connected layer that connects the 6D feature vectors into a layer with eight nodes. The output of this layer is reshaped into a 4×2 array. We pass the 4×2 array through a convolution layer with eight filters of size 3×3. The image is upsampled by two and passed through a convolution layer to form eight 8×4 images. The upsampling and convolution steps are repeated five more times increasing the size of the image by a factor of two at every step until we get eight 256×128 images. Each convolution layer uses *rectified linear unit*(ReLU) activation layers and a dropout of p=0.5. Finally the eight 256×128 image maps are combined using an additional CNN layer followed by a softmax activation, which converts the final output image into a probability map. Table I summarizes the architecture.

B. *Training*

The training of this network is challenging because (1) the network architecture is deep and diverging (i.e., instead of starting from several nodes and converging into few classes, our network starts with six nodes and ends with 32,768 labels (i.e. 256×128), and (2) unlike a regular multi-class classification problem, the labels need to be spatially consistent (i.e., the errors between classes are not equal). To address these challenges, we gradually add layers into the network by training the network after each cycle of upsampling. This approach is adopted instead of training the entire network at once. This process is possible, because we can quantize the continuous value into different levels of the discretion, and, hence, we have ground truth generated for each intermediate layer. To train the network, we add a CNN layer with a single channel output, followed by a softmax activation. The network has seven upsampling stages starting with an output image with 4×2 pixels, and finishing with an output image with 256×128 pixels. This method makes the training process more manageable, as it provides a good initial estimate for each new layer of upsampling. Notice that this approach increases the resolution of the predicted gaze region. The spatial dependency is implicitly considered by using CNN and gradual upsampling, as the pixels that are closer to each other share the nodes in the early layers. Figure 1 shows the outputs of the model at each stage of the training process.

We use Keras [26] with Tensorflow [27] backend to design and train our model. We change the learning rate and number of epochs depending on the training stage. The initial stage is trained for 1,000 epochs with a learning rate of 1e-3 to obtain a good initial model. Stages 2 - 5 are trained for 200 epochs with a learning rate of 1e-3. The last two stages are trained with a lower learning rate of 1e-4 for 500 epochs. We use Adam optimizer [28] in all the stages. The loss function is given by the KL-divergence between the one-hot coded target image and the predicted image.

IV. DATABASE

We use the naturalistic database presented in Jha and Busso [5] using the UTDrive vehicle. Drivers were asked to look at predefined markers multiple times while driving. We attached 21 markers inside the car, in the field of vision of the driver (e.g., windshield, mirrors, radio, and gear). Figure 2(a) shows some of the markers. We collected data from 16 drivers (10 males, 6 females). The data consists of three phases. In phase I, the subjects completed the task when the car was parked. In phase II, the subjects completed the task while driving the car. In phase III, the subjects were asked to turn their head completely toward each marker while the car was parked. This study only uses data collected in phase II. The drivers wore a headband with AprilTags [29], as shown in Figure 2(b). Since the accuracy of head pose estimation algorithms in a naturalistic driving environment is still a challenging problem [30], we rely on the head pose estimation obtained from the headband. While relying on the headband limits the deployment of the algorithm in real world applications, it provides more reliable training data to build our framework. The framework can be easily adapted when more robust head pose estimation algorithms become available.

To normalize the head pose values for each driver, the positions from the tags are averaged over a long-term driving data of each subject, and the value is subtracted from each frame. The rotation angles are also averaged in the quaternion space using *spherical linear interpolation* (slerp), and all the frames are rotated by the negative of the average. To calculate the gaze angles, we estimate the unit vector ($\hat{g}_{(x,y,z)}$) pointing to the target gaze location ($Gaze_{loc}$) from the head position ($H_{position}$):

$$\hat{g}_{(x,y,z)} = \frac{Gaze_{loc} - H_{position}}{\|Gaze_{loc} - H_{position}\|} \tag{1}$$

The horizontal angle θ is given by the angle between the projection of the gaze vector in the x-z plane and the z-axis. The vertical angle ϕ is given by the angle between the gaze vector and the x-z plane.

$$\begin{aligned} \theta_{gaze} &= \arctan \frac{g_x}{g_z} \\ \phi_{gaze} &= \arctan \frac{g_y}{\sqrt{g_x^2 - g_z^2}} \end{aligned} \tag{2}$$

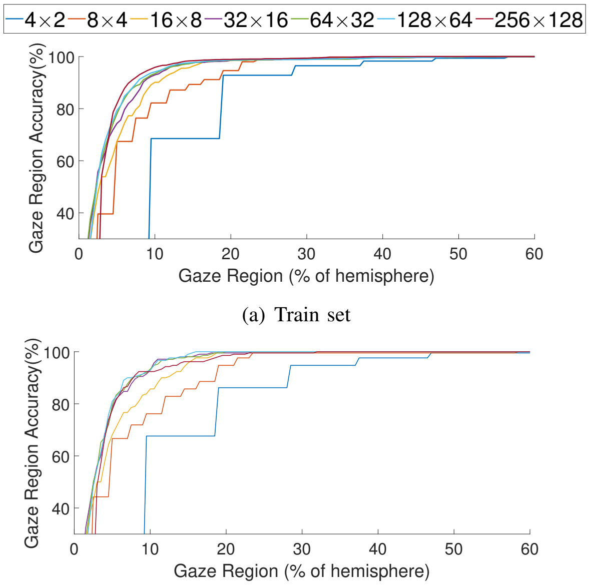
We test our model with subject independent partitions. The data from fourteen subjects are used to train the model, and data from two subjects are used to test the data. We repeat the process five times using different partitions. We report the average results over the five data partitions.

V. Results

We present experimental evaluations to study the effectiveness and reliability of the proposed method. Section V-A studies the prediction efficiency of the proposed method after every upsampling stage. We also compare the result of our model with the previously proposed method based on the GPR algorithm [6] (see Section II). Section V-B demonstrates the use of this model by visualizing two examples where the map is predicted from the head pose inputs.

A. Accuracy versus Precision

Our aim is to design a system that provides an accurate estimation of gaze within a region of confidence (saliency map). There is a trade-off between the size of the confidence region and the accuracy of the model (whether the region includes the true gaze). If the area is too large, the estimation does not carry enough information. If the area is too small, the accuracy may not be good for real world applications. Hence, we study the performance of the model by comparing the accuracy with the size of the region of confidence. The softmax output of the proposed framework creates a probability map. We can set a threshold to determine the size of the confidence region. It the threshold decreases, the gaze area increases, which results in higher accuracy but worse spatial resolution. We quantify the size of the confidence region as the percentage of the predicted region from half of a sphere centered on the driver with 180° of horizontal and 180° of vertical angles. This sphere approximates the driver's field of view. This metric does not depend on the radius of the sphere, which would be important if we want to project the region onto the windshield.

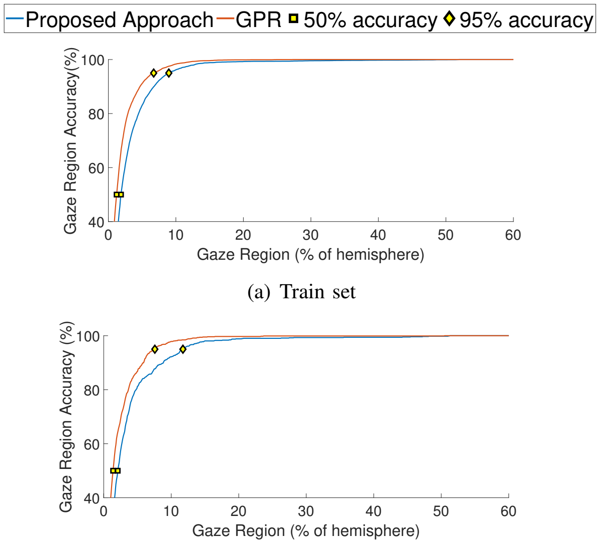


(a) Train set

(b) Test set

Fig. 3. Accuracy versus precision at different intermediate resolution for the proposed network.

Figure 3 plots the area of confidence against the percentage of the gaze included in the selected area at different resolution. The model is tested after each stage of upsampling. The x-axis provides the percentage of half of the sphere covered by the predicted gaze region. The y-axis provides the accuracy indicating whether the gaze direction is included in the predicted gaze region. With low resolution architectures (e.g., 4×2), the accuracy of the gaze region is high, but each interval represents a large area reducing the spatial resolution of the model. With each subsequent upsampling, the resolution increases but the confusion between adjacent labels also increases. Since we are interested in the gaze region, instead of the actual pixel, this is not a problem as the accuracy increases with respect to the predicted area. Increasing the resolution also increases the computational complexity and the time to train the model. Therefore, we stop at 256×128, which gives us a reasonable resolution (∼1°). We observe a small degradation in performance when



(a) Train set

(b) Test set

Fig. 4. Accuracy versus precision results for the proposed architecture and the GPR model [6]. The figures highlight the 50% and 95% accuracy points.

TABLE II

Percentage of the half of the sphere included in the predicted gaze region when the accuracy is 50%, 75%, and 95%.

	Proposed model		GPR	
	Training Data	Test Data	Training Data	Test Data
50%	**1.87**	**2.08**	**1.22**	**1.39**
75%	**3.75**	**3.95**	**2.44**	**3.06**
95%	**8.95**	**11.73**	**6.71**	**7.57**

evaluating the models on the test set.

Figure 4 compares the performance of the final output layer with 256×128 resolution with the model based on the GPR method proposed by Jha and Busso [6]. An advantage of the proposed method is that we can choose any arbitrary resolution depending on the requirements. In this study, we evaluate the performances obtained by the 256×128 model. Table II reports the percentage covered by the predicted region of half of the sphere when the accuracy is 50%, 75% and 95%. We observe that the results of the proposed model closely resemble the results of the GPR model. However, the accuracy of the GPR model is slightly higher. We hypothesize that one of the factor for the lower performance of the proposed model is the lack of resolution on the train data for the target gaze locations, as we only have 21 markers. The gaze prediction tends to converge to square regions centered at the location of these markers (see gaze region predicted in Figures 1 and 5). This problem can be addressed by using more markers, or a database with equipments to intrusively record the drivers gaze for each frame. Unlike the GPR model, the proposed framework is a non-parametric model, which has more potential as it can learn arbitrary distributions from the data.

B. Visualizing the Gaze Regions

We show two examples in Figure 5 to visualize the output predicted by the model. Figures 5(b) and 5(e) show the predicted gaze region when the accuracy is 50% for the head poses described in Figures 5(a) and 5(d). These regions correspond to 1.74% of the half sphere. The green dots indicate the target gaze location. Likewise, Figures 5(c), and 5(f) show the gaze region when the accuracy is 95%. These regions correspond to 9.82% of the half sphere. The values range between -110° and +110° in yaw angles and between -100° and +10° in pitch angles. In Figures 5(a)-5(c) the ground truth gaze location is marker 2. Our model predicts with high precision that the gaze is directed towards the front. Similarly, in Figures 5(d)-5(c), the subjects head is turned toward the left. In the predicted map, we observe that the model predicts with high probability that the subject is looking at the left of the windshield. We observe that the true gaze direction is close to the predicted gaze region with 50% accuracy. The region for 95% accuracy included the target gaze direction.

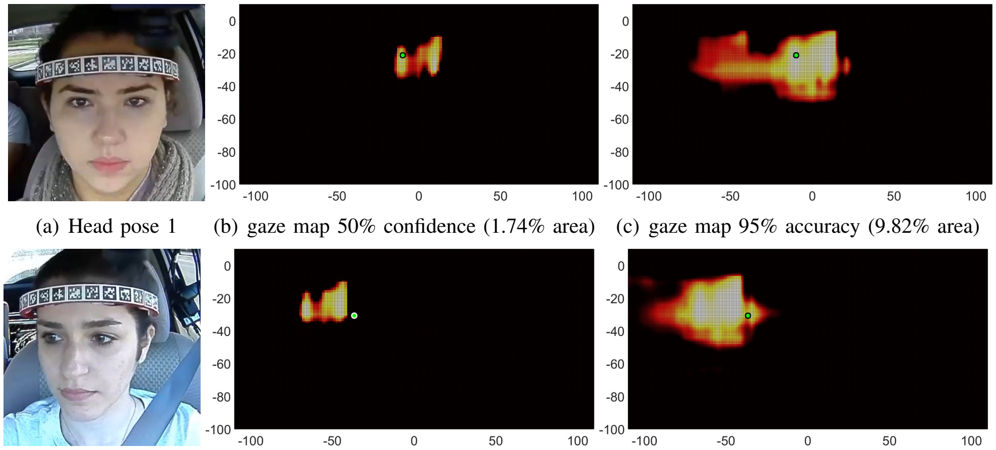
VI. Conclusions

This study proposed an efficient method to estimate the driver's visual attention using a generic probability distribution using deep learning. We gradually up-sample the resolution of the gaze region, which increases the accuracy and spatial resolution of the prediction. We enforced that nodes have strong correlation with their neighbor nodes using CNNs. The intensity of each pixel in the image indicates the probability of the target gaze to be directed toward a given direction. The proposed architecture is a novel solution for probabilistic regression, leveraging the predicted power of dense networks. This model provides an appealing solution, not only for our target application, but also for other predictive tasks. This architecture can be easily adapted to other fields to formulate regression problems as classification tasks. For example, if the number of variables is greater than two, the approach can be implemented with multi-dimensional convolution layers.

While the proposed approach is a non-parametric model without making any assumption, we observe that the performance is slightly lower than the GPR model. To address this problem, we can investigate more advanced models. We can add multiple CNNs in each layer to make the interpolation function more generic. We can also use more advanced loss functions, which are sensitive to the distance between nodes. Adding eye related features is also expected to improve the accuracy of the system, increasing the number of inputs, which in the current model is only six. The dataset that we used has gaze labels only at fixed points. This is an important limitation of our implementation. If the training data include higher resolution of the target gazes, the model will learn more accurate gaze regions of arbitrary shapes. Databases collected with gaze tracker can provide us with richer continuous labels to train the model.

References

[1] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep. CMU-CS-94-102, January 1994.



(a) Head pose 1 (b) gaze map 50% confidence (1.74% area) (c) gaze map 95% accuracy (9.82% area)

(d) Head pose 2 (e) gaze map 50% confidence (1.74% area) (f) gaze map 95% confidence (9.82% area)

Fig. 5. Gaze region predicted by our models for two examples. The green dots show the location of the target gaze.

[2] K. Mora and J. Odobez, "Person independent 3D gaze estimation from remote RGB-D cameras," in *IEEE International Conference on Image Processing (ICIP 2013)*, Melbourne, Australia, September 2013, pp. 2787–2791.

[3] E. Murphy-Chutorian and M. Trivedi, "HyHOPE: Hybrid head orientation and position estimation for vision-based driver head tracking," in *IEEE Intelligent Vehicles Symposium (IV 2008)*, Eindhoven, The Netherlands, June 2008, pp. 512–517.

[4] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Portsmouth, NH: ACM, October 2012, pp. 149–154.

[5] 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)*, Rio de Janeiro, Brazil, November 2016, pp. 2157–2162.

[6] ——, "Probabilistic estimation of the driver's gaze from head orientation and position," in *IEEE International Conference on Intelligent Transportation (ITSC)*, Yokohama, Japan, October 2017, pp. 1630–1635.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org, 2016, pp. 1747–1756.

[8] A. Tawari and M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *IEEE Intelligent Vehicles Symposium (IV 2014)*, Dearborn, MI, June 2014, pp. 344–349.

[9] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. Los Angeles, CA, USA: IEEE, June 2017, pp. 849–854.

[10] 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*, vol. 12, no. 1, pp. 254–267, March 2011.

[11] Y. Liang and J. Lee, "Combining cognitive and visual distraction: Less than the sum of its parts," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 881–890, May 2010.

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

[13] ——, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, April 2016.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *ArXiv e-prints (arXiv:1609.03499)*, vol. abs/1609.03499, September 2016.

[15] C. M. Bishop, "Mixture density networks," Aston University, Tech. Rep., February 1994. [Online]. Available: http://www.ncrg.aston.ac.uk/

[16] L. Torgo and J. Gama, "Regression by classification," in *Brazilian symposium on artificial intelligence*. Springer, 1996, pp. 51–60.

[17] E. Frank and M. Hall, "A simple approach to ordinal classification," in *European Conference on Machine Learning*. Springer, 2001, pp. 145–156.

[18] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," *Interspeech, 2017 (to apear)*, 2017.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[21] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2018–2025.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[23] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[25] C. Hettinger, T. Christensen, B. Ehlert, J. Humpherys, T. Jarvis, and S. Wade, "Forward thinking: Building and training neural networks one layer at a time," *arXiv preprint arXiv:1706.02480*, 2017.

[26] F. Chollet, "Keras: Deep learning library for Theano and TensorFlow," https://keras.io/, April 2017. [Online]. Available: https://github.com/fchollet/keras

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation (OSDI 2016)*, Savannah, GA, USA, November 2016, pp. 265–283.

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

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

[30] 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)*, Yokohama, Japan, October 2017, pp. 1624–1629.