Real-time Eye Gaze Direction Classification Using Convolutional Neural Network

Anjith George, *Member, IEEE,* and Aurobinda Routray, *Member, IEEE*

***Abstract*—Estimation eye gaze direction is useful in various human-computer interaction tasks. Knowledge of gaze direction can give valuable information regarding users point of attention. Certain patterns of eye movements known as eye accessing cues are reported to be related to the cognitive processes in the human brain. We propose a real-time framework for the classification of eye gaze direction and estimation of eye accessing cues. In the first stage, the algorithm detects faces using a modified version of the Viola-Jones algorithm. A rough eye region is obtained using geometric relations and facial landmarks. The eye region obtained is used in the subsequent stage to classify the eye gaze direction. A convolutional neural network is employed in this work for the classification of eye gaze direction. The proposed algorithm was tested on Eye Chimera database and found to outperform state of the art methods. The computational complexity of the algorithm is very less in the testing phase. The algorithm achieved an average frame rate of 24 fps in the desktop environment.**

***Index Terms*—Eye gaze tracking, Convolutional Neural Network, EAC, eye tracking, Gaze direction estimation.**

I. Introduction

Understanding human emotions and cognitive states are essential in developing a natural human-computer interaction system (HCI). Systems which can identify the affective and cognitive states of humans can make the interaction more natural. The knowledge of mental processes can help computer systems to interact intelligently with humans. Recently, many works have been reported [1], [2] investigating the use of facial expression in HCI. Human eyes provide rich information about human cognitive processes and emotions. The movement patterns of eyes contains information about fatigue [3], diseases [4], etc. Pupil dilation has also been used as an indicator to study cognitive processes [5]. The nature of eye movements is unique for each. Recently, several works has been proposed to use eye movement pattern as a biometric [6], [7].

Most of the works related to facial expression are constrained to desktop environments. However, with the new development of wearable devices like Google Glass [8] and other augmented reality goggles there are more opportunities for using eye analysis techniques for understanding the affective and cognitive states of the users.

The patterns in which the eyes move when humans access their memories is known as eye accessing cues (EAC). The patterns in this non-visual gaze directions have been reported to contain information regarding mental processes. In Neuro-Linguistic Programming (NLP) theory [9], eye accessing cues gives information about the mental processes from the direction of eye gaze. These movements are reported to be related to the neural pathways which deal with memory and sensory information. The direction of the iris in the socket can give information regarding various cognitive processes. Each direction of non-visual gaze is associated with different cognitive processes. The meanings of the different EACs are shown in Fig 1. More details about EAC model can be found from [9]. Even though the EAC theory is not 100 % accurate, recent studies [10], [11] have found correlation which encourages further research in the field. A critical review of EAC method can be found in [12].

Information retrieval systems can work in a better way if the context is known. Knowledge of the cognitive states can be useful in providing the context in HCI.

Most of the approaches for gaze estimation uses active infrared based methods which require expensive hardware [13]. In this paper, we develop a real-time framework which can detect eye gaze direction using off-the-shelf, low-cost cameras in desktops and other smart devices. Estimation of gaze location from webcam often requires cumbersome calibration procedure [14]. We treat the gaze direction classification as a multi-class classification problem, avoiding the need for calibration. The eye directions obtained can be used to find the EAC and thereby infer the user's cognitive process. The information obtained can be useful in the analysis of interrogation videos, human-computer interaction, information retrieval, etc.

The highlights of the paper are shown below:

- Proposes a real-time framework for eye gaze direction classification
- We use a Convolutional Neural Network based gaze direction classifier, which is robust against eye localization errors
- The proposed approach outperforms state of the art algorithms in gaze direction classification
- The algorithm achieves an average frame rate of 24 fps in desktop environment

II. Related works

There are many works related to gaze tracking in desktop environments, an excellent review of the methods can be found in [15]. In this section, we limit the discussion to the recent state of the art works related to eye gaze direction estimation and EAC.

Vrânceanu *et al*. proposed a method [16] for automatic classification of EAC. The information from color space is

This is a preprint of the accepted paper. To appear in IEEE International Conference on Signal Processing and Communication, SPCOM 2016.

used in their approach. The relative position of iris and sclera in the eye bounding box is used to classify the visual accessing cues. Vrânceanu *et al*. proposed another method [17] for finding EAC using iris center detection and facial landmark detection. They used isophote curvature based method for iris center localization. The relative position of iris center is used with the fiducial points for a better estimate of eye gaze direction. In [18] Radlak *et al.* presented a method for gaze direction estimation method in static images. They used an ellipse detector with a Support Vector based verifier. The bounding box is obtained using the hybrid projection functions [19]. Finally, the gaze direction is classified using Support Vector Machine (SVM) and random forests.

Recently Vrânceanu *et al*. [20] proposed another approach for eye direction detection using component separation. Iris, sclera, and skin are segmented and the features obtained are used in a machine learning framework for classifying the eye gaze direction. Zhang *et al*. [21] applied convolutional neural network for gaze estimation. They combined the data from face pose estimator and eye region using a CNN model. They have trained a regression model in the output layer.

In most of the related works, the general framework is by using three cascaded stages. Face detection, eye localization, and classification. The localization or classification errors in any of the cascaded stages will result in the reduction of overall accuracy. The computational complexity of the methods is another bottleneck. In this work, we aim at increasing the accuracy of eye gaze direction classification. The developed algorithm is robust against noise, blur, and localization errors. The computational complexity is less in the testing phase, and the proposed algorithm achieves an average 24 fps in a PC based implementation. The proposed framework is described in the following section.

III. Proposed Algorithm

The overall framework proposed is shown in Fig. 2. Different stages of the algorithm are described below.

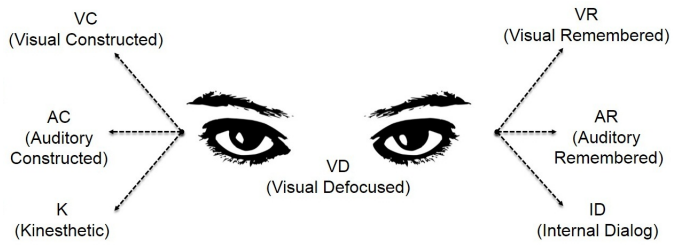

Fig. 1: Different EACs in NLP theory

A. Face detection and eye region localization

The first stage in the algorithm is face detection. We have used a modified version of Viola-Jones method [22] in this paper. The method used is fast and invariant to in-plane rotations, the accuracy and trade offs can be found in [23]. Once the face region is localized, next stage is to obtain the eye region. We have used two different methods for obtaining the

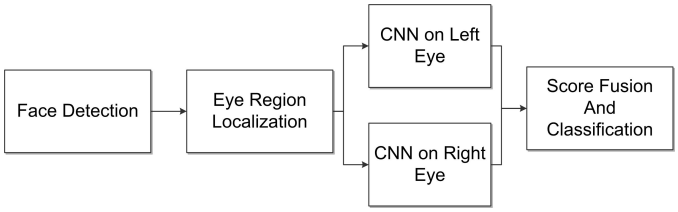

Fig. 2: Schematic of the overall framework

eye region. In the first method, the eye region for classification is obtained geometrically from the face bounding box returned by the face detector (ROI). The dimension of the eye region is shown on an image from HPEG database [24] in Fig. 3. The eye regions obtained are re-scaled to a resolution of 42×50 for the subsequent stages (ROI). In the second method, we used a facial landmark detector to find the eye corners and other fiducial points.

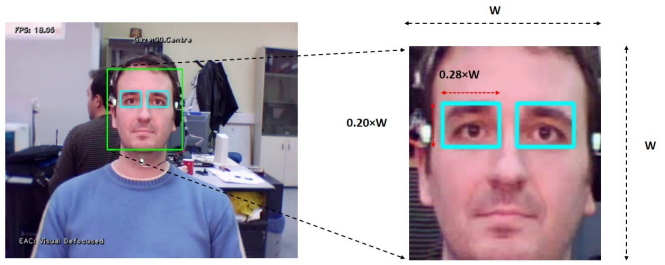

Fig. 3: Eye region localization using geometric approach (ROI)

1) Facial landmark localization: Localization of facial landmarks helps in constraining the eye region for classification. Ensemble of Randomized Tree approach (ERT) [25] approach is used for localizing facial landmarks. The face bounding box obtained from the preceding stage is used as the input to the algorithm. The locations of the facial landmarks are regressed using a sparse subset of pixels from the face region. The algorithm is very fast and works even with partial labels. The details of the algorithm can be found in [25]. The eye corner location returned from the landmark detector is used to select the eye region used in the subsequent classification stage.

B. Eye gaze direction classification

The eye region obtained from the previous stage is used in a multiclass classification framework for predicting the EAC classes. Convolutional Neural Network (CNN) is used for the classification. The details of the model used are described below.

1) Convolutional Neural Network (CNN): The convolutional neural network represents a type of feed-forward neural network which can be used for a variety of machine learning tasks. Krizhevsky *at al*. [26] used a large CNN model for the classification of images in imagenet database. Even though the training time is huge, the accuracy and robustness of CNNs are better than most of the standard machine learning

algorithms. In our approach, we have used a CNN model with three convolution stages. The input stage consists of images of dimension 42×50 (or 25×15 in the case of ERT). In the first convolutional layer, 24 filters of dimension 7×7 are used. This stage was followed by a rectifier linear unit (ReLU). ReLU layer introduces a non-linearity to the activations. The non-linearity function can be represented as:

$$f(x) = \max(0, x) \tag{1}$$

where, x is the input and $f(x)$ the output after the ReLU unit. A max pooling layer is added after the ReLU stage. Max pooling layer performs a spatial sub-sampling of each output images. We have used 2×2 max-pooling layers which reduce the spatial resolution to half. Two similar stages with filter dimensions 5×5 and 3×3 are also added. After the convolutional, ReLU, and max-pooling layers in the third convolutional layer, the outputs from all the activations are joined in a fully connected layer. The number of output nodes corresponds to the number of classes in the particular application. The structure of the network is shown in Fig. 4. The softmax loss is used over classes as the error measure. Cross entropy loss is minimized in the training.

The cross entropy loss (L) is defined as:

$$L\left(f(x), y\right) = -y \ln(f(x)) - (1-y)\ln(1 - fx(x)) \tag{2}$$

where x is the vector to be classified, $y \in \{0, 1\}$, where y is the label

The cross entropy loss is convex and can be minimized using Stochastic Gradient Descent (SGD) [27] algorithm. The size of convolution kernels remains same for both ERT and ROI (7×7, 5×5 and 3×3).

2) Classification of eye gaze direction: Two CNN networks are trained independently for left and right eyes. The scores from both the networks are used to obtain the class labels.

$$score = \left(\frac{score_L + score_R}{2}\right) \tag{3}$$

where, $score_L$ and $score_R$ denote the scores obtained from left and right CNNs respectively.
The class can be found out as the label with maximum probability as:

$$class = \arg\max_{label}(score) \tag{4}$$

IV. Experiments

We have conducted experiments in Eye Chimera database [28], [29] which contains all the seven EAC classes. This dataset contains images of 40 subjects. For each subject, images with different gaze directions (for various EAC classes) are available. The total number of images in the dataset is 1170. The ground truth for class labels and fiducial points are also available.

A. Evaluation procedure

The database was randomly split into two equal proportions. Training and testing are performed on two completely disjoint 50% subsets to avoid over-fitting. CNN require a large amount of data in the training phase for better results. The size of the database is relatively small. We have used data augmentation in the training set images to solve this issues. Rotations, blurring and scaling, are performed in the images in the training subset to increase the number of training samples. Two CNNs were trained separately for left and right eye. In the testing phase, the scores from both left and right eye CNN models are combined to obtain the label of the test image.

We have considered both 7 class and 3 class classification in this work. The methodology followed is same in both the cases.

B. Results

The results obtained from the experiments in Still Eye Chimera dataset is shown here. The classification accuracy was high in 3 class scenario compared to the accuracy in 7 class case.

We have conducted experiments with the two different methods proposed. In the first case, the eye region localization is carried out using geometrical relations. Explicit landmark detection is avoided in this case. This method is denoted as ROI. This approach reduces one stage in the overall framework. Additionally, the robustness of the algorithm against localization errors can be tested. In the second algorithm, we use the ERT based landmark detection scheme. The eye corners obtained are used to constrain the region for subsequent classification stage. The region obtained in each image is resized to a resolution of 20×15 for further processing.

In both the cases (ROI and ERT), the data was divided into two 50% subsets. CNNs were trained separately for left and right eyes using data augmentation. Testing was done 50% disjoint testing set to avoid over-fitting effects. All the experiments were repeated in both 3 class and 7 class scenario. In 3 class case, we use only classes left, center and right.

The results obtained by using only one eye are shown in Table I.

Combining the information from both eyes improves the accuracy. The results obtained using both the eyes and the comparison with the state of the art is shown in Table III.

In both the cases, the proposed method outperforms all the state of the art algorithms in eye gaze direction classification. Highest accuracy is obtained with ERT+CNN algorithm. The individual accuracies achieved in the 7 class case is shown in Table II.

The confusion matrix for 3 class and 7 class case (ERT+CNN) are shown in Fig. 5 and Fig. 6.

C. Discussion

The proposed algorithm outperforms all the state of the art results reported in the literature. From the confusion matrix, it can be seen that most of the mix classifications are in differentiating between right, down right, etc. The classification accuracy is poor in the vertical direction (As observed in [20]). This can be attributed to the lack of spatial resolution in the vertical direction. Most of the cases iris is partly occluded by eyelids in extreme corners. This makes it difficult to classify them accurately. With the larger amount of

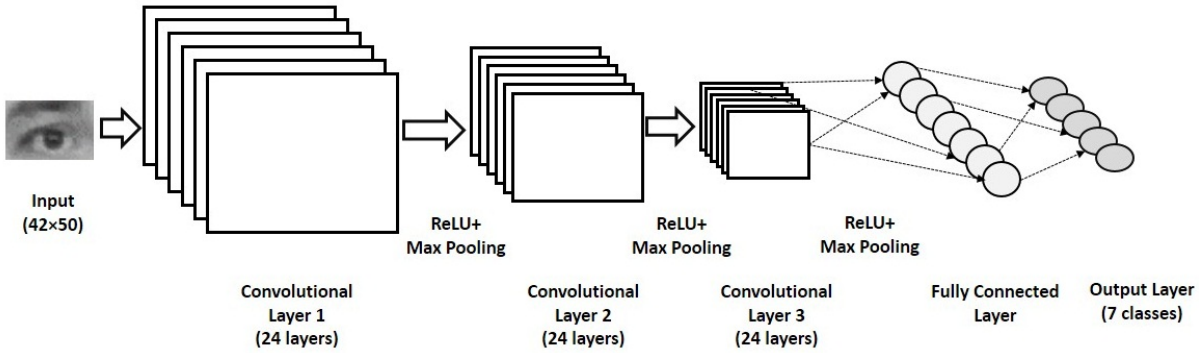

Fig. 4: Architecture of the CNN used

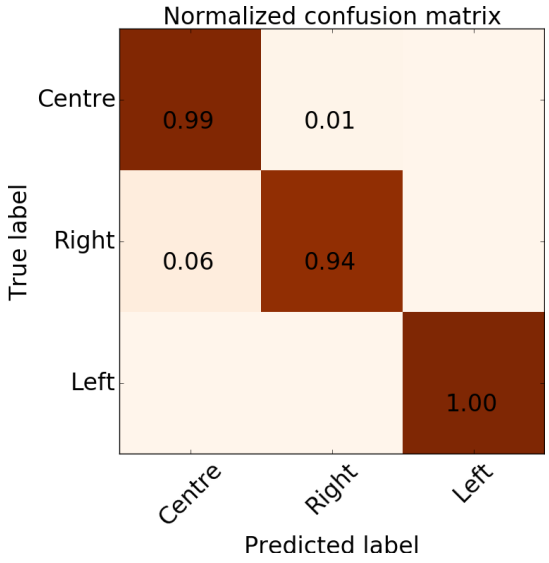

Fig. 5: Confusion matrix for 3 classes (ERT+CNN)

TABLE I: Comparison of accuracy of classification using only one eye

Eye Boundingbox Localization method	Eye direction classification Method	Recognition Rate 7 class (%)	Recognition Rate 3 class (%)
BoRMaN [30]	Valenti [31]	32.00	33.12
Zhu [32]	Zhu [32]	39.21	45.57
Vrânceanu [20]	Vrânceanu [20]	77.54	89.92
Proposed (Geometric)	**Proposed (CNN)**	**81.37**	**95.98**
Proposed (ERT)	**Proposed (CNN)**	**86.81**	**96.98**

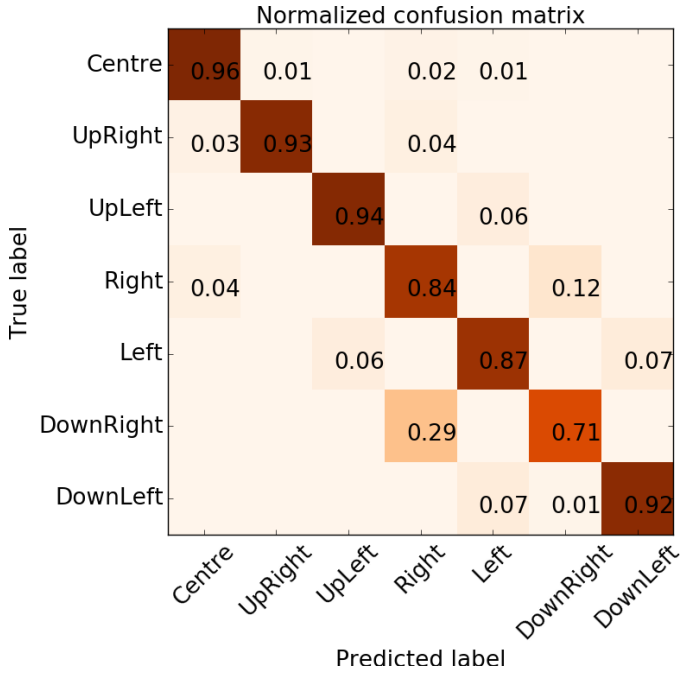

Fig. 6: Confusion matrix for 7 classes (ERT+CNN)

TABLE II: Accuracy in classification of each class (%)

Method	VD	VR	VC	AR	AC	ID	K
Proposed (ROI)	96	79	87	77	79	75	93
Proposed (ERT)	97	93	94	84	87	71	91

labeled data, the algorithm could perform even better. Using the color information in the CNN can improve the accuracy even further. A temporal filtering of the predicted labels can improve the accuracy in the case of video.

V. Conclusion

In this work, a framework for real-time classification of eye gaze direction is presented. The estimated eye gaze direction is used to infer eye accessing cues, giving information about the cognitive states. The computational complexity is very less; we achieved frame rates around 24 Hz in Python implementation in a 2.0 GHz Core i5 PC running Ubuntu 64 bit (4GB RAM). The per-frame computational time is 42 ms, which is much less than that of the other state of the art methods (250 ms in [20]). Off the shelf webcams can be used for computing the Eye gaze direction. The proposed algorithm works even with in-plane rotations of the face. The eye gaze direction obtained can also be used for human-computer interaction applications. The

TABLE III: Accuracy in EAC classification (%) when both the eyes are used

Dataset	Classes	[31] + [30]	Zhu [32]	[20]	**Proposed (ROI)**	**Proposed (ERT)**
Still Eye	7	39.83	43.29	83.08	**85.58**	**89.81**
Chimera	3	55.73	63.01	95.21	**97.65**	**98.32**

computational complexity of the algorithm in testing phase is less, which makes it suitable for smart devices with low-resolution cameras using pre-trained models.

VI. Acknowledgements

The authors would like to thank Dr. Corneliu Florea for providing the database.

References

[1] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[2] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.

[3] L. L. Di Stasi, R. Renner, A. Catena, J. J. Cañas, B. M. Velichkovsky, and S. Pannasch, "Towards a driver fatigue test based on the saccadic main sequence: A partial validation by subjective report data," *Transportation research part C: emerging technologies*, vol. 21, no. 1, pp. 122–133, 2012.

[4] Y. Terao, H. Fukuda, A. Yugeta, O. Hikosaka, Y. Nomura, M. Segawa, R. Hanajima, S. Tsuji, and Y. Ugawa, "Initiation and inhibitory control of saccades with the progression of parkinson's disease–changes in three major drives converging on the superior colliculus," *Neuropsychologia*, vol. 49, no. 7, pp. 1794–1806, 2011.

[5] S. D. Goldinger and M. H. Papesh, "Pupil dilation reflects the creation and retrieval of memories," *Current Directions in Psychological Science*, vol. 21, no. 2, pp. 90–95, 2012.

[6] A. George and A. Routray, "A score level fusion method for eye movement biometrics," *Pattern Recognition Letters*, 2015.

[7] C. D. Holland and O. V. Komogortsev, "Complex eye movement pattern biometrics: Analyzing fixations and saccades," in *Biometrics (ICB), 2013 International Conference on*. IEEE, 2013, pp. 1–8.

[8] T. Starner, "Project glass: An extension of the self," *Pervasive Computing, IEEE*, vol. 12, no. 2, pp. 14–16, 2013.

[9] R. Bandler and J. Grinder, "Frogs into princes: Neuro linguistic programming. 1979."

[10] J. Sturt, S. Ali, W. Robertson, D. Metcalfe, A. Grove, C. Bourne, and C. Bridle, "Neurolinguistic programming: a systematic review of the effects on health outcomes," *British Journal of General Practice*, vol. 62, no. 604, pp. e757–e764, 2012.

[11] R. Vranceanu, L. Florea, and C. Florea, "A computer vision approach for the eye accesing cue model used in neuro-linguistic programming," *Sci. Bull. Univ. Politehnica Bucharest Ser. C*, vol. 75, no. 4, pp. 79–90, 2013.

[12] G. Diamantopoulos, S. I. Woolley, and M. Spann, "A critical review of past research into the neuro-linguistic programming eye-accessing cues model," *Current Research in*, p. 8, 2009.

[13] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer Science & Business Media, 2007, vol. 373.

[14] A. George and A. Routray, "Fast and accurate algorithm for eye localisation for gaze tracking in low resolution images," *IET Computer Vision*, 2016.

[15] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 3, pp. 478–500, 2010.

[16] R. Vrânceanu, C. Vertan, R. Condorovici, L. Florea, and C. Florea, "A fast method for detecting eye accessing cues used in neuro-linguistic programming," in *Intelligent Computer Communication and Processing (ICCP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 225–229.

[17] R. Vranceanu, C. Florea, L. Florea, and C. Vertan, "Automatic detection of gaze direction for nlp applications," in *Signals, Circuits and Systems (ISSCS), 2013 International Symposium on*. IEEE, 2013, pp. 1–4.

[18] K. Radlak, M. Kawulok, B. Smolka, and N. Radlak, "Gaze direction estimation from static images," in *Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on*. IEEE, 2014, pp. 1–4.

[19] F. Song, X. Tan, S. Chen, and Z.-H. Zhou, "A literature survey on robust and efficient eye localization in real-life scenarios," *Pattern Recognition*, vol. 46, no. 12, pp. 3157–3173, 2013.

[20] R. Vrânceanu, C. Florea, L. Florea, and C. Vertan, "Gaze direction estimation by component separation for recognition of eye accessing cues," *Machine Vision and Applications*, vol. 26, no. 2-3, pp. 267–278, 2015.

[21] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2015.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[23] A. Dasgupta, A. George, S. Happy, and A. Routray, "A vision-based system for monitoring the loss of attention in automotive drivers," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 4, pp. 1825–1838, 2013.

[24] S. Asteriadis, D. Soufleros, K. Karpouzis, and S. Kollias, "A natural head pose and eye gaze dataset," in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. ACM, 2009, p. 1.

[25] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1867–1874.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[28] L. Florea, C. Florea, R. Vrânceanu, and C. Vertan, "Can your eyes tell me how you think? a gaze directed estimation of the mental activity," in *Proceedings of the British Machine Vision Conference. BMVA Press*, 2013, pp. 60–1.

[29] R. Vrânceanu, C. Florea, L. Florea, and C. Vertan, "Nlp eac recognition by component separation in the eye region," in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 225–232.

[30] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2729–2736.

[31] R. Valenti and T. Gevers, "Accurate eye center location and tracking using isophote curvature," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[32] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.