Deep neural network ensemble architecture for eye movements classification

Marko Arsenovic, Srdjan Sladojevic, Darko Stefanovic, Andras Anderla
Department of Industrial Engineering and Management
Faculty of Technical Sciences
Novi Sad, Serbia
arsenovic@uns.ac.rs, sladojevic@uns.ac.rs, andras@uns.ac.rs, darkoste@uns.ac.rs

Abstract— **Up to now, eye tracking technologies have been used for different purposes in various industries, from medical to gaming. Eye tracking methods could include predicting fixations, gaze mapping or movement classification. Recent advances in deep learning techniques provide possibilities for solving many computer vision tasks with high accuracy. Authors of this paper propose a novel deep learning based architecture for eye movement classification task. Proposed architecture is an ensemble approach which employs deep convolutional neural networks that run in parallel, for both eyes separately, for visual feature extractions along with recurrent layers for temporal information gathering. Dataset images for training and validation were gathered from standard web camera and pre-processed automatically using dedicated tools. Overall accuracy of developed classifier on the validation set was 92%. Proposed architecture uses relatively small networks which brings the possibility of real time usage (successfully tested on 15-20fps) on regular CPU. Classifier achieved overall accuracy of 88% on the real-time test, using standard laptop and web camera.**

Keywords-data deep learning; eye tracking; image classification; time-series prediction; convolutional networks, recurrent networks

I. INTRODUCTION

Eye tracking represents technique of eye position recording and movement. With the constant improvement in both, software and hardware technologies, eye tracking brought interest to many researchers and it has been applied in various industries. Over the last years, eye tracking methods were used in market, medical, psychology, gaming, UI/UX (user interface/user experience) and robotic research.

By applying eye tracking tools, many leading brands evaluate their advertisements, designs and customers' shopping behavior measuring their attention and navigation or using gaze mapping. Similar eye tracking methods became popular also in IT industry for evaluating websites' UX by measuring attention and hot spots. Medical researchers use these methods for diagnosing diseases such as OCD (Obsessive Compulsive Disorder) or Parkinson's and Alzheimer's disease [1, 2, 3]. Eye movement classification found important application in gaming and HCI (human-computer interaction) [4]. In these types of eye tracking implementations, people are able to control hardware devices by only moving their eyes.

In the last decade, deep learning methods proved to be very efficient in computer vision problems. Deep convolutional neural networks became the most successful in many benchmarks for image classification and detection problems [5, 6, 7]. Based on those results, authors of this paper propose ensemble deep neural network architecture for eye movement classification. The proposed architecture employs deep CNNs for visual features and recurrent network for temporal features for solving eye movement classification task with high accuracy.

The rest of the paper is organized as follows: Section II presents related work, Section III presents methods of developing proposed eye tracking classifier, Section IV presents achieved results and discussion related to them, and Section V holds the conclusion.

II. RELATED WORK

In many computer vision tasks deep learning approach replaced traditional algorithms and outperformed them in many aspects. In recent years, along with standard machine learning methods, numerous researchers embraced these new approaches in solving eye tracking tasks.

Rello et al. in [8] applied previous eye tracking measurements to build a classifier to detect readers with dyslexia. They used common machine learning algorithm, support vector machine (SVM) for classification task reaching final accuracy of 80.18% with 10-fold cross validation.

Wang et al. in [9] proposed regression based deep convolutional neural network (RCNN) for feature learning for predicting eye fixations. Using this approach, they introduced highly accurate 2D regression-based and 3D model-based eye gaze tracking methods.

Hoppe et al. in [10] employed deep CNN as an end-to-end eye movement detector from the continuous gaze data stream. Their method included two main steps. First step transforms window holding input raw gaze data into frequency domain by applying the FFT (fast Fourier transform), while in second step that encoded input is processed by CNN that outputs the predicted eye movement class. The proposed method outperformed traditional machine learning approaches achieving 74% for multi class prediction.

978-1-5386-4907-7/18/$31.00 ©2018 IEEE

Krafka et al in [11] introduced novel architecture iTracker, CNN specialized for eye tracking which is able to run in real time (10-15 fps) on a modern mobile device. Input of the iTracker CNN include right and left eye along with the face images detected and cropped from the original video frame while output of the model is the gaze prediction. Proposed architecture consists of three same CNNs training in parallel for left and right eye and cropped face, which are merged with fully connected layers along with the face grid (position of face in the original image). Stacking layers in CNNs applied in iTracker architecture are very similar to AlexNet [12]. Along with novel architecture, in the same paper they introduced the first large-scale eye tracking dataset, GazeCapture, containing data from over 1450 people with 2.5M frames. This new large-scale dataset was used to train and evaluate iTracker network.

Motivated by the results of deep neural networks in various eye tracking tasks, authors of this paper propose novel deep ensemble architecture for eye movement classification task. Entire procedure of the proposed method is described in detail in further sections.

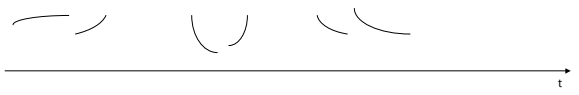
III. MATERIALS AND METHODS

The complete method of creating the eye tracking classifier is described further in detail. Process of building eye movement classifier is divided into several phases, including dataset gathering along with pre-processing, designing network architecture, training and validating the proposed model.

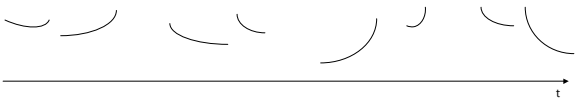
A. Dataset

Custom dataset was used for building the eye movements tracking classifier proposed in this paper. Building dataset included several necessary steps: capturing images of face using simple web camera, detecting eyes in the frame, creating new images containing both eyes separately, image pre-processing and storing them labelled in the file system. Eye movement represents time-series classification, which requires creating time window containing the frames of eyes' images of the movement process. Every window holds entire movement captured in time. Specialized Python script using OpenCV [13] library was developed for dataset gathering process.

First step of dataset gathering process was recording every eye movement separately. For this experiment two different eye movements were used, V and Z shape movements along with the idle state, as it is represented by Figure 1.



a) V sign movement with eyes in time



b) Z sign movement with eyes in time

c) Idle state of eyes in time

Figure 1. Three classes of eye movements in time: V, Z and Idle

In Figure 1. it can be observed irises' movements in time to create V shape (left side corner, down and right side corner), Z shape (left top corner, right top corner, left down corner and right down corner) and idle state with small movements with central fixations.

Every eye movement was recorded 500 times with 50 frames per recording and was stored in the file system. Second step of the dataset preparation process included finding eyes in every frame and cropping them seperatly applying HAAR Cascades [14].

Third step was pre-processing of the images by reshaping them to 32x32, and grayscaling for faster training time by significantly reducing the number of weights. Histogram normalization [15] was also applied in the pre-processing stage in order to extract shadows from the images. After pre-processing stage, dataset consisted of 500 labeled window samples per class for both eyes seperatly where every window contained 50 time-series images.

Fugure 2. represents original grayscaled and pre-processed eye images.

a) original grayscale

b) pre-processed

Figure 2. Original grayscaled and pre-processed eye images

B. Neural Network Architecture

Eye movement classification model needs to be able to extract both visual and time-dependent features. In order to achieve that, authors of this paper proposed deep learning based ensemble method combining CNN networks for visual tasks and recurrent GRU (Gated Recurrent Units) for time based prediction. Proposed architecture is displayed in Figure 3.

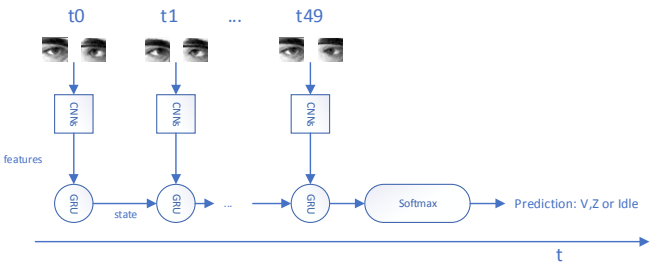


Figure 3. Neural Networks Ensembled Architecture

Visual features are extracted by employing two 3-layered CNNs in parallel, one for each eye seperatly. These CNNs are merged through the last fully connected layer which extract important information from pairs of eyes. Figure 4.

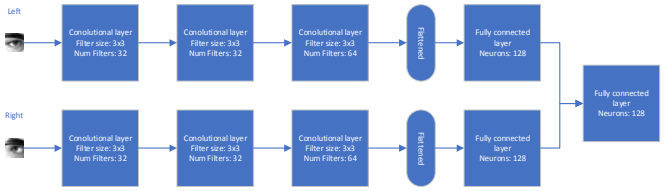


Figure 4. CNN Architecture

When visual information is extracted, ensemble network exploit GRU recurrent layer to update current state with the information based on a previous state and current time step. When the 50th image from the window is processed, GRU's state represents input of a softmax classifier where output represents the class of the eyes' movements. Tensorflow [16], deep learning framework, was used to build and evaluate the proposed deep ensemble architecture.

IV. Results and Discussion

Deep ensemble neural network described in previous section was trained for 1000 epochs on 450 window samples from the dataset, while 50 window samples are used as a validation set. Final overall accuracy of the model was 92%. Confusion matrix with prediction on validation set is displayed in Table 1.

TABLE I. Confusion Matrix for the Validation Dataset

Classified as			
Z	*V*	***Idle***	
13	3	1	Z
1	16	0	V
0	2	14	Idle

Due to the fact that deep neural networks reduce overfitting and improve general classification accuracy when trained with larger datasets, overall accuracy of the proposed model could be improved by adding more recording windows to the dataset or by augmenting the existing data.

Visual feature extraction could be enhanced by adding more learning layers to the CNN or by applying transfer learning using pre-trained models of state-of-the-art deep CNN architectures such as Inception or ResNet [17, 18].

Authors of this paper chose the proposed CNN architecture with the grayscale images in order to reduce number of learning weights which resulted with smaller computation time. According to that, the proposed system could work in real time (15-20fps) using todays average CPU.

The proposed classifier was also tested in real time situation using regular webcam and laptop computer. Overall accuracy of 88% was slightly less comparing to the controlled environment test. Confusion matrix is displayed in Table 2.

TABLE II. Confusion Matrix for the Test Dataset

Classified as			
Z	*V*	***Idle***	
12	2	0	Z
2	12	1	V
0	1	13	Idle

From the observation of the captured frames in real-time testing and false positives, it could be seen that some of the frames were blurred due to the small head movements while camera recording, which resulted in false predictions. That sort of false prediction could be potentially corrected by introducing certain augmentation methods in the pre-processing stage, which could adopt the CNN for certain distortion and noise in the captured frames.

V. Conclusion

In the recent years, there was massive expansion of deep learning technologies usage in real world applications. Eye tracking tasks have significant role in many industries from helping people with disabilities to getting insights in psychology and improving market revenue.

In this paper, new deep neural network ensemble architecture is introduced for solving eye movement classification task with high accuracy but with modest hardware capabilities. This lightweight architecture could be applied in real time, which makes it suitable for practical applications.

High classification accuracy motivates authors of this paper to pursue the further research in this area. The future work could include expanding the dataset with more classes and data while exploring different CNN architectures for visual tasks, also evaluate different types of recurrent layers for temporal features. In some cases, deep convolutional neural networks proved very useful even in time-series predictions. Changing recurrent layer with convolutional ones could be evaluated in eye movement classification tasks.

REFERENCES

[1] Stuart, S., et al. "Accuracy and re-test reliability of mobile eye-tracking in Parkinson's disease and older adults." Medical engineering & physics 38.3 (2016): 308-315.

[2] Crawford, Trevor J., et al. "The disengagement of visual attention in Alzheimer's disease: a longitudinal eye-tracking study." Frontiers in aging neuroscience 7 (2015).

[3] Bradley, Maria C., et al. "Obsessive–compulsive symptoms and attentional bias: An eye-tracking methodology." Journal of behavior therapy and experimental psychiatry 50 (2016): 303-308.

[4] Majaranta, Päivi, and Andreas Bulling. "Eye tracking and eye-based human–computer interaction." Advances in physiological computing. Springer London, 2014. 39-65.

[5] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.

[6] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[7] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.

[8] Rello, Luz, and Miguel Ballesteros. "Detecting readers with dyslexia using machine learning with eye tracking measures." Proceedings of the 12th Web for All Conference. ACM, 2015.

[9] Wang, Kang, Shen Wang, and Qiang Ji. "Deep eye fixation map learning for calibration-free eye gaze tracking." Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. ACM, 2016.

[10] Hoppe, Sabrina, and Andreas Bulling. "End-to-end eye movement detection using convolutional neural networks." arXiv preprint arXiv:1609.02452 (2016).

[11] Krafka, Kyle, et al. "Eye tracking for everyone." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[13] Bradski, Gary, and Adrian Kaehler. Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.", 2008.

[14] Wilson, Phillip Ian, and John Fernandez. "Facial feature detection using Haar classifiers." Journal of Computing Sciences in Colleges 21.4 (2006): 127-133.

[15] Soha, J. M., and A. A. Schwartz. "Multispectral histogram normalization contrast enhancement." 5th Canadian Symposium on Remote Sensing. 1979.

[16] Abadi, Martín, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).

[17] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition.2015

[18] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.