

Eye Centre Localisation with Convolutional Neural Networks in High- and Low-Resolution Images

Wenhao Zhang(✉) and Melvyn L. Smith

Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England, Bristol BS16 1QY, UK
wenhao.zhang@uwe.ac.uk

Abstract. Eye centre localisation is critical to eye tracking systems of various forms and with applications in variety of disciplines. An active eye tracking approach can achieve a high accuracy by leveraging active illumination to gain an enhanced contrast of the pupil to its neighbourhood area. While this approach is commonly adopted by commercial eye trackers, a dependency on IR lights can drastically increase system complexity and cost, and can limit its range of tracking, while reducing system usability. This paper investigates into a passive eye centre localisation approach, based on a single camera, utilising convolutional neural networks. A number of model architectures were experimented with, including the Inception-v3, NASNet, MobileNetV2, and EfficientNetV2. An accuracy of 99.34% with a 0.05 normalised error was achieved on the BioID dataset, which outperformed four other state-of-the-art methods in comparison. A means to further improve this performance on high-resolution data was proposed; and it was validated on a high-resolution dataset containing 12,381 one-megapixel images. When assessed in a typical eye tracking scenario, an average eye tracking error of 0.87% was reported, comparable to that of a much more expensive commercial eye tracker.

Keywords: Eye centre localisation · Eye tracking · Deep learning

1 Introduction

Eye tracking has seen a long history of extensive use in differing application domains since its early-stage development in psychological studies tracking eye movements in reading [1, 2]. The increasing variety and quantity of commercial eye trackers [3, 4] and related research works [5] have provided strong evidence that this technology bears high potentials in contributing to multi-disciplinary research and to assistance with day-to-day human activities. For example, eye tracking was employed by marketers and designers to measure the effectiveness of advertisements in magazines [6]; and more recently to understand how users would view websites [7]. When applied to a distinctive discipline, eye tracking could assist with diagnosis of neurodegenerative diseases by providing eye movement biomarkers able to assess cognitive performance [8]. Another trending area of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13375, pp. 373–384, 2022.
https://doi.org/10.1007/978-3-031-10522-7_26

application concerns human-computer interaction, in the form of gaming-based learning [9] or control of a computer [10], for example.

Eye centre localisation is fundamental to development of eye tracking technologies, and it has a significant impact on eye tracking accuracy and precision. Choice of approach to eye centre localisation also directly affects complexity, cost, and usability of an eye tracking system. Generally, systems which work at a close proximity to eyes such as head-mounted eye trackers, and which have complex designs such as those utilising active illumination, can offer greater precision and accuracy but have lower affordability and usability. For example, a typical head-mounted system employed by psychological or medical research [11] can cost tens of thousands of pounds or more, resulting in limited adoption of such technologies. More importantly, uncomfortable intrusiveness of a head-mounted device is likely to undermine exhibition of natural user behaviours which are often critical to such studies. Although remote eye trackers exist, they usually have a limited range of working distance (e.g. 40 cm) as well as a lower accuracy than their head-mounted counterparts [12]. While offering a less expensive option, they are still far from being widely affordable.

Motivated by these challenges, this paper presents a deep learning based approach for eye centre localisation with a single camera, which aims to accelerate development of a cost-effective solution for remote and passive eye tracking with a high precision and accuracy. This research makes a contribution by investigating the effectiveness of different convolutional neural network (CNN) architectures for eye centre localisation, and by assessing the impact of image resolution on localisation accuracy. Based on research findings, a possible means to further improve the performance of the proposed approach is discussed and future works are recommended.

2 Related Works

Depending on whether active illumination is required, an eye centre localisation approach can be categorised as either active or passive, with the former being the predominant category due to the number of benefits active illumination can offer. Commonly, active illumination takes the form of infrared (IR) or near infrared light oriented towards the eyes. When an IR source is positioned close to the optical axis of a camera, active illumination causes the pupil to be lit, creating a brighter elliptical shape in high contrast to its neighbourhood. This is known as the bright pupil method. Different to this setup, the dark pupil method requires off-axis positioning of IR light(s), leading to a darker pupil but also an enhanced contrast [13]. In both methods, active illumination also causes cornea reflections, which provide additional features to allow conversion from eye centre locations to gaze positions taking into consideration head movements. For example, the Tobii Pro Nano device [12] employs an approach that switches between the bright pupil mode and the dark pupil mode according to environmental conditions. However, active eye centre localisation methods are often faced with real-world challenges. For example, as sunlight has a broad IR spectra, its interference to active illumination could lead to inaccurate eye centre localisation results in an outdoor application [14].

Different to these active methods, passive eye centre localisation approaches do not rely on active illumination, but they employ inherent eye appearance and/or geometry

features under ambient lighting. While this can lead to a reduced system cost and complexity, there is a higher demand for and emphasis on overcoming interfering features, such as eyelids and makeup, which will often appear to be more prominent in unstructured image data due to a lack of active illumination. This has motivated various studies to exploit geometric features of the eyes by modelling circular (but realistically elliptical) contours of the pupil and iris. For example, one of the approaches [15] designed an objective function of gradient based features, drawing on the fact that gradients residing on the edge of a disk would be oriented towards its centre. This approach achieved reasonable eye centre localisation results by solving this optimisation problem, although its performance would be largely compromised by presence of strong gradients from shadows, other facial features such as eyebrows, and occluded pupil or iris. To deal with these interfering features, a study [16] based on image topography explicitly detected eyebrows such that false candidates from these regions could be removed during a multi-resolution analysis of iris features. Another method [17] tried to tackle this challenge by employing a two-stage approach to perform a coarse-to-fine estimation. In this approach, a convolution operator would be used to obtain an initial estimation of the eye centres based on geometric features; boundary tracing and ellipse fitting were then used to refine previous estimations. Similarly, the method proposed by [18] designed a two-stage approach combining gradient features and isophote features filtered by a bespoke selective oriented gradient filter to progressively reduce interfering features at a global scale before carrying out a local-level analysis in order to achieve improved accuracy. However, specularities and shadows mimicking geometric characteristics of pupil and iris still pose a fundamental challenge.

As machine learning approaches generally excel in dealing with complex patterns that cannot be easily and explicitly characterised, their capabilities have been leveraged to improve the accuracy of eye centre localisation. A method proposed in [19] employed a number of deep neural networks for face detection, eye detection, and openness assessment in succession; all as preliminary stages to eye centre localisation. However, instead of taking advantage of machine learning throughout all stages, they designed heuristic-based features for analysis of the iris. Similarly, the method presented in [20] only employed machine learning in a preliminary stage that served to identify a smaller region of interest. To achieve this, the Dlib toolkit [21] was utilised to detect facial landmarks including those of eye corners and eyelids. Both approaches reported incremental improvements, but occluded pupil and iris, as well as presence of glasses, still caused a large localisation error. A different approach [22] embracing a higher utilisation of machine learning designed an end-to-end CNN for predicting eye centre locations within face regions. Building on established network architectures, such as Inception-v3 [23] and ResNet [24], this approach achieved a significantly higher eye centre localisation accuracy.

3 Methodology

3.1 Datasets and Data Capture Experiments

Most state-of-the-art methods have reported results on a publicly available dataset, namely the BioID dataset [25, 26]. Therefore, we incorporated this dataset for developing and validating the proposed method to facilitate quantitative and comparative performance evaluation. This dataset consists of 1521 grayscale images of 23 different subjects, and it contains realistic challenges such as a variable ambient illumination, presence of glasses, and other types of inter-subject and intra-subject eye appearance variability. The images are of a low resolution, i.e., 384 × 288 pixels. Inclusion of a large background area means that the effective region of a face or an eye in these images has even fewer pixels. While deep learning is known to be able to handle small images effectively in a wide range of challenges such as object detection and classification, low-resolution data will inevitably limit eye tracking performance. In view of this, we constructed a new dataset that has a higher image resolution, and we assessed its impact on CNN models for eye centre localisation. To capture these data, we recruited ten participants at the University of the West of England (UWE), Bristol, including seven males and three females, aged between 18 and 55 years and of different ethnic backgrounds. The research experiments had been approved by UWE Faculty Research Ethics Committee (reference No: HAS.19.08.017). All data were recorded in a laboratory environment where overhead lighting was consistent but could lead to self-shadowing on faces. Each participant watched a five-minute video on a 27-inch screen positioned at approximately 60 cm away showing a moving visual target used to trigger eye movements. A chin rest was used to restrict large head movements such that eye movements would become more prominent. A machine vision camera (FLIR Grasshopper3) beneath the screen, angled towards the face regions, was set to capture images at 160 frames per second at a size of 1024 by 1024 pixels. A workstation with an Intel Xeon E5-2630 processor and 128 GB RAM hosted the control programme and interfaced with the camera. The experiment was repeated four times for each participant (i.e., four trials). The system setup and experiment procedure we employed are representative of those in eye tracking based psychological and medical studies, such as [27] and [28]. This would help place our research in context and would facilitate a more meaningful performance validation.

We manually annotated a random subset of data (12,381 images) from the first trials only, with pixel coordinates of eyes centres. The resolution of our data is 10 times as high as the BioID dataset (considering pixel count of an image). The camera field-of-view is primarily covered by face regions, leaving little background. Therefore, the pixel count of a face region is effectively over 20 times higher than the BioID dataset. The 60 cm screen-to-participant distance, the 27-in. screen size, and the moving visual target also caused eyes to move to extreme positions, horizontally, vertically, and diagonally. This introduced various levels of iris occlusions (by eyelids) contributing to data variability.

3.2 CNNs for Eye Centre Localisation

Building on the success of our prior investigation [22], in this study, we further employed different CNN architectures to solve eye centre localisation as a regression problem. We utilised an established CNN backbone to learn various levels of features. A global averaging pooling layer then condenses these features before passing them through fully connected layers, finally outputting the left and right eye centre coordinates. The Mean Squared Error (MSE) was used as the loss function. For the BioID dataset, we employed the Dlib library for face detection which successfully removed the background in each image. The face regions then became the input to our CNN models. For the new dataset we constructed, as the area of background is negligible, we directly input raw images to the CNN models. A number of backbones were experimented with, including Inception-v3 [23], the NASNet [29], MobileNetV2 [30] and the EfficientNetV2 (V2-L and V2-B3) [31]. Without unfolding the backbones and exhaustively listing all their layers, the general CNN architecture can be illustrated in Fig. 1.

Backbone	Inception-v3	
Concatenate	Input:	(8, 8, 320)
		(8, 8, 768)
		(8, 8, 768)
		(8, 8, 192)
	Output:	(8, 8, 2048)
+		
Global average pooling	Input:	(8, 8, 2048)
	Output:	2048
+		
Fully connected	Input:	2048
	Output:	4

Fig. 1. An illustration of the CNN architecture, using Inception-v3 as an example

All these different backbone architectures achieved outstanding top-1 and top-5 accuracy on the ImageNet [32] validation dataset, which evidenced their exceptional capabilities of feature extraction in general image classification tasks. Despite this, one of the problems these models may encounter is loss of spatial precision in a landmark localisation task. This is caused by spatial pooling of features typically occurring in a CNN, which is intended to progressively enlarge the overall receptive field while reducing the number of parameters. While it can serve this dual purpose, it effectively downsamples feature maps passed through the network; and therefore, this reduces localisation capability of the network. In view of this, without modifying the network architecture, we investigated into the impact of image resolution on CNN performance. In addition, from the perspective of eye tracking, acquisition of high-resolution images can also contribute to a higher accuracy.

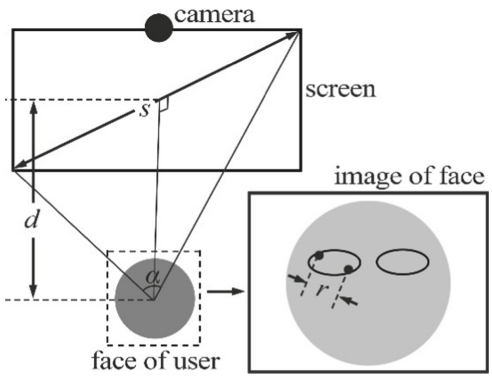

Fig. 2. Illustration of a screen-based eye tracking scenario where the two dots within image of face represent the extreme pupil centre positions, giving a maximum displacement value of r pixels.

Taking a typical eye tracking scenario for example (depicted in Fig. 2) and assuming that a user gazes at information on a screen with a diagonal size of s cm, while a remote eye tracker performs image-based eye centre localisation at a distance of d cm; an eye centre localisation error of one pixel, namely the smallest value in the discrete image domain, corresponds to an error of e_{deg} degrees that can be calculated by Eq. (1).

$$e_{deg} = \frac{2arctan\left(\frac{s}{2d}\right)}{r} \cdot \frac{180}{\pi} \tag{1}$$

This means that, when user distance d and screen size s are fixed, an increase in image resolution will proportionally increase displacement of pupil centre r, leading to a reduced error. Although high-resolution images can potentially reduce eye tracking error, an increased number of input neurons will drastically increase computation, consequently demanding a much larger GPU memory. Therefore, we employed the Dlib toolkit to detect eye corners in each face image, which informed extraction of two square eye regions from each face image, such that these smaller regions of interest could then become model input. This allowed utilisation of a reasonable batch size for training without having to downsample input size. Consequently, the model was changed to output coordinates of a single eye centre at a time.

4 Results and Discussions

The CNN models were implemented with Python 3.8, Tensorflow and Keras 2.8.0, and were evaluated on a workstation with an Intel i9-9940X CPU, 64 GB memory, and a Nvidia Titan RTX GPU.

As the number of samples in the BioID dataset is not particularly large in the context of deep learning, we employed the five-fold cross validation method with a 8:2 split for training and validation in each fold. We followed the commonly used relative error metric proposed by [23] to evaluate eye centre localisation accuracy. This metric calculates an

error as the Euclidean distance between eye centre estimates and their ground truth before normalising this distance relatively to the pupillary distance. This is formulated by Eq. (2).

$$e_{max} = \frac{\max(d_l,\ d_r)}{\omega} \tag{2}$$

where d_l and d_r are the absolute errors for the eye pair, and ω is the pupillary distance in pixels. The maximum value of d_l and d_r after normalisation is defined as the maximum normalised error e_{max}. A normalised error of 0.05 would be equivalent of the average pupil diameter. To train the model, cropped face regions from images were used, as mentioned previously. They were resized to 299 × 299 pixels, meaning that they had to be significantly upscaled. We used the Adam optimiser, a batch size of 32, a learning rate of 0.001. We observed that the model would generally converge within 300 epochs without showing severe overfitting; therefore, we do not emphasise the importance of hyperparameter tuning here. We first show in Fig. 3 that this CNN approach, when employing different backbones, could achieve highly accurate eye centre localisation results.

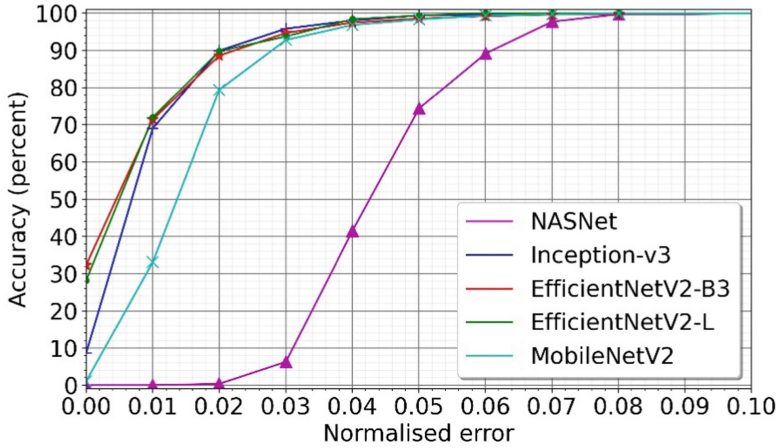

Fig. 3. Accuracy curves of the CNN model on the BioID dataset when utilising different architectures

These accuracy curves can demonstrate that accuracies with a tolerance of 0.05 normalised error approached 100%. The highest accuracy overall was obtained by EfficientNetV2-B3 that was one of the most compact models experimented with, able to perform inference at 46ms per step. The EfficientNetV2-L model achieved a similar overall accuracy, but its parameter count is over seven times higher. The Inception-v3 model produced a slightly worse accuracy but it could perform inference faster at 36ms per step. The lowest accuracy was from NASNet despite its large model size and topological depth, which overfitted when having not received further hyperparameter tuning. A few examples of inaccurate localisation results (where $e_{max} > 0.05$) by the

EfficientNetV2-B3 model are shown in Fig. 4. It can be seen that, in a few instances, the CNN predictions appear to be more accurate than the ground truth. Admittedly, manual annotation of data is prone to error, but this could be overcome by a CNN given a sufficient amount of correctly labelled data. We then compared the CNN performance with that of four other state-of-the-art methods on the BioID dataset. The results are shown comparatively in Table 1.

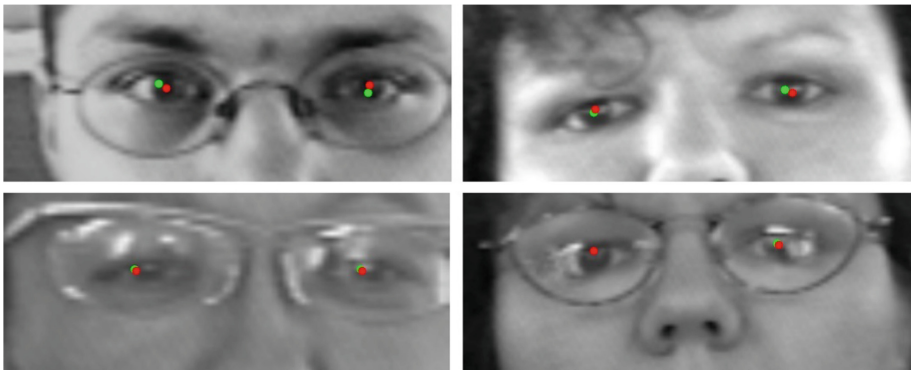

Fig. 4. Examples of inaccurately localised eye centres where the green dots represent the ground truth and red dots the predictions. Only the eye regions are displayed to give the areas of interest a better visibility. (Color figure online)

Table 1. Comparative evaluation of the proposed method on the BioID dataset. The * notation means that the exact value of accuracy has not been provided by the referenced works, but it was approximated according to the accuracy curves.

Method	Normalised error			
	$e_{max} \leq 0.03$	$e_{max} \leq 0.05$	$e_{max} \leq 0.10$	$e_{max} \leq 0.25$
EfficientNetV2-L	93.71%	99.34%	100%	100%
EfficientNetV2-B3	94.65%	98.52%	99.92%	99.92%
Inception-v3	95.80%	99.34%	99.92%	99.92%
MobileNetV2	92.76%	98.27%	99.92%	99.92%
NASNet	6.25%	74.34%	100%	100%
[20]	52%*	94.50%	100%	100%
[19]	/	94.25%	98.40%	99.45%
[17]	50%*	85.08%	94.30%	98.13%
[16]	62%*	85%*	94%*	99.5%*

As this demonstrates superior performance of the proposed method, we followed a similar process to evaluate it further on the high-resolution dataset we constructed, utilising the Inception-v3 model. As this dataset has approximately ten times as many

images as the BioID dataset, it was partitioned into training, validation, and testing with a split of 8:1:1. With an input image size of 512 × 512 pixels, the model arrived at the minimum validation loss at epoch 62 within five hours of training. When further trained beyond this point, the model did not severely overfit. More details can be seen in Fig. 5.

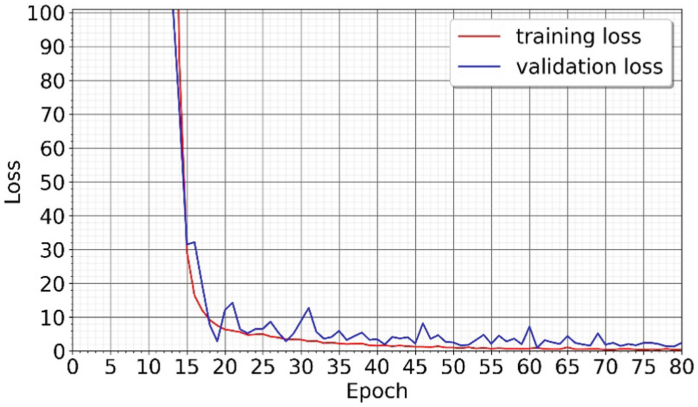

Fig. 5. Training and validation loss curves on the high-resolution dataset and the Inception-v3 model

Model performance was then evaluated on input images downsampled to different sizes, respectively. As mentioned in the preceding section, eye regions cropped by Dlib were also used to train the model. We compared these results and report them in Fig. 6.

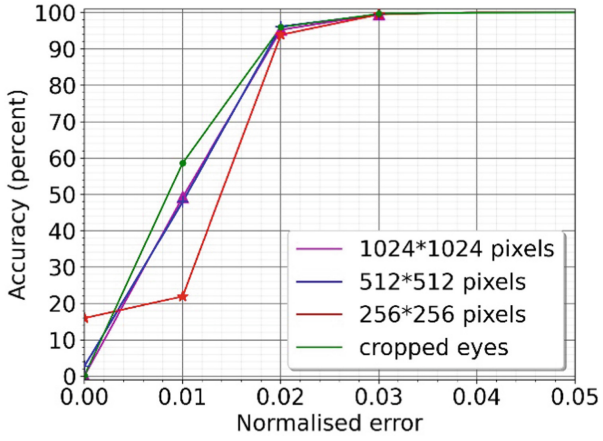

Fig. 6. Accuracy curves on the high-resolution dataset

When model input received uncropped images, high-resolution images contributed to improved model performances. When cropped eye regions were used to train the model, a further improvement on overall accuracy was achieved, as shown in Fig. 6.

This is one of the ways to increase the pixel count of the regions of interest (i.e., the eyes) without dramatically increasing the size of model input. Examples of inaccurate eye centre localisation were shown in Fig. 7. A normalised error $e_{max} > 0.02$ was used, as all instances had an error smaller than 0.05.



Fig. 7. Examples of inaccurate eye centre localization on the high-resolution dataset

Following Eq. (1), the average eye localisation error was calculated to be 0.87° and was similar for different resolutions of input. We also experimented with data augmentation to simulate data variability caused by head movements, such as roll, yaw and pitch of the head, horizontal and vertical translation, and a variable distance to camera. This was achieved by applying a moderate rotation and translation to batches of images during training. In all cases, we observed that data augmentation did not improve model performance on the datasets. The reasons are likely that the training data employed already sufficiently modelled data variability such as eye occlusion, shadows and specularities in the testing set; and also that the CNN models were able to differentiate between useful features from these interferences. However, we argue that when the models are exposed to a large number of unseen faces and significant head movements, data augmentation will likely contribute to an improved model performance.

5 Conclusions

This study investigated into CNN capabilities for eye centre localisation on high- and low-resolution images. Based on the low-resolution BioID dataset, it first validated the performance of a number of state-of-the-art CNN architectures, including Inception-v3, NASNet, MobileNetV2, and EfficientNetV2; and demonstrated superior localisation accuracies in comparison to other similar approaches. Both the Inception-v3 model and the EfficientNetV2 model achieved an eye localisation accuracy close to 100% with a normalised error of 0.05. Following this, the CNN model was evaluated on a high-resolution dataset; and it showed that a higher resolution could improve eye centre

localisation performance, given the same model architecture. Additionally, we used Dlib to detect and crop eye regions as a preliminary step to CNN based eye centre localisation. By removing a large amount of background, this effectively increased resolution of the regions of interest only, such that the size of input to model remained relatively small. The proposed eye centre localisation approach, when placed in a typical eye tracking scenario, could achieve an error of 0.87°, comparable to a much more expensive commercial eye tracker. We also experimented with data augmentation for improving data variability, which did not lead to a higher model performance. However, we argue that the proposed data augmentation technique is likely to be able to make a contribute when large head movements are present.

In our future works, we will investigate into model architectures optimised for receiving high-resolution images as a means to further reduce eye centre localisation error. We also intend to combine localisation with tracking to overcome inaccuracies caused by eye blinks as well as to improve efficiency. We will also employ this cost-effective eye tracking approach to facilitate healthcare studies such as early diagnosis of eye diseases and neurodegeneration.

References

1. Rayner, K.: Eye guidance in reading: fixation locations within words. Perception **8**(1), 21–30 (1979)
2. McConkie, G.W., Rayner, K.: The span of the effective stimulus during a fixation in reading. Percept. Psychophys. **17**(6), 578–586 (1975)
3. Tobii pro Homepage. https://www.tobiipro.com/. Accessed 02 Apr 2022
4. Gazepoint Homepage. https://www.gazept.com/. Accessed 08 Mar 2022
5. Duchowski, A.T.: A breadth-first survey of eye-tracking applications. Behav. Res. Methods Instrum. Comput. **34**(4), 455–470 (2002)
6. Krugman, D.M., Fox, R.J., Fletcher, J.E., Rojas, T.H.: Do adolescents attend to warnings in cigarette advertising? An eye-tracking approach. J. Advert. Res. **34**(6), 39–53 (1994)
7. Hervet, G., Guérard, K., Tremblay, S., Chtourou, M.S.: Is banner blindness genuine? Eye tracking internet text advertising. Appl. Cogn. Psychol. **25**(5), 708–716 (2011)
8. Crawford, T.J., Devereaux, A., Higham, S., Kelly, C.: The disengagement of visual attention in Alzheimer's disease: a longitudinal eye-tracking study. Front. Aging Neurosci. **7**, 118 (2015)
9. Kiili, K., Ketamo, H., Kickmeier-Rust, M.D.: Evaluating the usefulness of eye tracking in game-based learning. Int. J. Ser Games **1**(2), 51–65 (2014)
10. Zhang, X., Liu, X., Yuan, S.M., Lin, S.F.: Eye tracking based control system for natural human-computer interaction. Comput. Intell. Neurosci. **2017**, 1–9 (2017)
11. Mele, M.L., Federici, S.: Gaze and eye-tracking solutions for psychological research. Cogn. Process. **13**(1), 261–265 (2012)
12. Tobii pro nano. https://www.tobiipro.com/product-listing/nano/. Accessed 08 Mar 2022
13. Gneo, M., Schmid, M., Conforto, S., D'Alessio, T.: A free geometry model-independent neural eye-gaze tracking system. J. Neuroeng. Rehabil. **9**(1), 1–15 (2012)
14. Binaee, K., Sinnott, C., Capurro, K.J., MacNeilage, P., Lescroart, M.D.: Pupil Tracking under direct sunlight. In: ACM Symposium on Eye Tracking Research and Applications, pp. 1–4. Association for Computing Machinery, New York (2021)
15. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. Visapp **11**, 125–130 (2011)

16. Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., Cabeza, R.: Hybrid method based on topography for robust detection of iris center and eye corners. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **9**(4), 1–20 (2013)
17. George, A., Routray, A.: Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images. IET Comput. Vision **10**(7), 660–669 (2016)
18. Zhang, W., Smith, M.L., Smith, L.N., Farooq, A.: Gender and gaze gesture recognition for human-computer interaction. Comput. Vis. Image Underst. **149**, 32–50 (2016)
19. Ahmad, N., Yadav, K.S., Ahmed, M., Laskar, R.H., Hossain, A.: An integrated approach for eye centre localization using deep networks and rectangular-intensity-gradient technique. J. King Saud Univ.-Comput. Inf. Sci. (2022)
20. Khan, W., Hussain, A., Kuru, K., Al-Askar, H.: Pupil localisation and eye centre estimation using machine learning and computer vision. Sensors **20**(13), 3785 (2020)
21. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
22. Zhang, W., Smith, M.: Eye centre localisation with convolutional neural network based regression. In 2019 IEEE 4th International Conference on Image, Vision and Computing, pp. 88–94. IEEE (2019)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826. IEEE (2016)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE (2016)
25. The BioID. Face database (2014). https://www.bioid.com/About/BioID-Face-Database. Accessed 05 Feb 2019
26. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the Hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) Audio- and Video-Based Biometric Person Authentication. AVBPA 2001. Lecture Notes in Computer Science, vol. 2091. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45344-X_14
27. Crutcher, M.D., Calhoun-Haney, R., Manzanares, C.M., Lah, J.J., Levey, A.I., Zola, S.M.: Eye tracking during a visual paired comparison task as a predictor of early dementia. Am. J. Alzheimer's Dis. Other Dement.® **24**(3), 258–266 (2009)
28. Oyama, A., et al.: Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. Sci. Rep. **9**(1), 1–9 (2019)
29. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710. IEEE (2018)
30. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. IEEE (2018)
31. Tan, M., Le, Q.: Efficientnetv2: smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106. PMLR (2021)
32. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)