IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 12, 2017, accepted September 10, 2017, date of publication September 19, 2017,
date of current version October 12, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2754299

Webcam-Based Eye Movement Analysis Using CNN

CHUNNING MENG[1] AND XUEPENG ZHAO[2]
[1]Department of Electronic Technology, China Maritime Police Academy, Ningbo 315801, China
[2]Nankai University, Tianjin 300071, China

Corresponding author: Chunning Meng (mengchunning123@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61401105.

ABSTRACT Due to its low price, webcam has become one of the most promising sensors with the rapid development of computer vision. However, the accuracies of eye tracking and eye movement analysis are largely limited by the quality of the webcam videos. To solve this issue, a novel eye movement analysis model is proposed based on five eye feature points rather than a single point (such as the iris center). First, a single convolutional neural network (CNN) is trained for eye feature point detection, and five eye feature points are detected for obtaining more useful eye movement information. Subsequently, six types of original time-varying eye movement signals can be constructed by feature points of each frame, which can reduce the dependency of the iris center in low quality videos. Finally, behaviors-CNN can be trained by the time-varying eye movement signals for recognizing different eye movement patterns, which is capable of avoiding the influence of errors from the basic eye movement type detection and artificial eye movement feature construction. To validate the performance, a webcam-based visual activity data set was constructed, which contained almost 0.5 million frames collected from 38 subjects. The experimental results on this database have demonstrated that the proposed model can obtain promising results for natural and convenient eye movement-based applications.

INDEX TERMS Convolutional neural network, eye movement analysis, feature point detection, visual activity recognition.

I. INTRODUCTION

For over a hundred years, recording and analyzing eye movement has been investigated, where the methods evolve from artificial observational method, mechanical method, electrical method to optical method (computer vision based method) [1]. More recently, computer vision based eye movement has become a dominant method in several domains including cognitive science [2], psychology [3], medical diagnoses [4], market research [5], identity authentication [6] and eye-based human-computer interaction [7]. In particular, computer vision based methods have the advantages of lower invasive, higher precision and easier operation. Traditionally, eye movement based applications are implemented based on eye tracking and eye movement analysis, and eye movement analysis is implemented based on gaze estimation and detection of basic eye-movement types.

According to a survey conducted by Hansen [8], the taxonomy of the existing gaze estimation consists of feature-based and appearance-based methods. The feature-based method is the most popular gaze estimation method, which requires to extract gaze related local features. Feature-based methods can be divided into corneal reflection and shape-based methods, where the categorization is based on the adoption of external light sources [9]. Corneal-reflection methods that use multiple cameras and multiple infrared lights have been successfully applied such as Tobii and SMI eye tracker. The precision of this method is satisfactory since the pupil center and the glint can be easily extracted to calibrate the errors caused by head movements. However, complex calibration, high cost and ill-health caused by IR lights are unavoidable. Moreover, IR light based systems are not reliable when used in outdoor conditions [8]. Shape-based methods cannot obtain estimation with high accuracy [10], [11] and require high image quality with precisely extracted pupil, iris and eyelid edges [12], [13]. However, both the pupil and glint are often unavailable in videos captured by webcam. In general, the appearance-based methods [14], [15] do not require calibration of cameras and geometry data. The image content is used as input for estimating the underlying function for gaze points or gaze orientation. Although this method can

be more flexible, it is very sensitive to head movements, where the accuracy is limited. More recently, deep learning has become a promising tool for computer vision applications, by achieving remarkable performance gains [16]. By using CNN, superior performance of gaze estimation can be obtained to learn representations from huge amounts of data. Zhang *et al.*[9] proposed a method for in-the-wild appearance-based gaze estimation using multimodal CNN, where a dataset that contains 213,659 images is collected. Krafka *et al.* [17] developed an end-to-end eye tracking solution targeting mobile devices, and a large-scale dataset with almost 2.5 million frames was introduced to train the CNN models, which can achieve a significant reduction in errors. Due to the use of deep and multimodal networks and large-scale training dataset, these methods can be applicable in the unconstrained daily-life setting with arbitrary head poses. However, the estimation precision of these methods is not good enough for eye-based human-computer interaction. If a distance of 70cm is kept from the person to the computer screen, the error is around 7.7cm for Zhang's method. The error of GazeCapture around 1.5cm is also larger than the distance between two adjacent app icons on mobile phones.

Precise gaze estimation is a powerfull guarantee for the technology of eye movement analysis. Although deep convolutional nets have achieved breakthroughs for gaze estimation in low-cost and daily-life scenarios, the precision is still unsatisfactory. It should be noted that gaze point or orientation is not necessary for eye movement analysis in various domains such as nystagmus or dyslexia diagnoses, eye based lie detection, fatigue detection and eye-based activity recognition. A mapping function should exist between a length of eye movement videos and one type of activity or state. Time-varying eye movement information can be used for analyzing eye movement without considering gaze and related calibration [18]. The advantages of abandoning gaze estimation based model can be summarized as follows. (1) The error of gazing mapping calibration can be avoided. (2) More features related to eye movement (not just the center of iris which is unreliable due to eyelid occlusion) can be employed to compensate the error of iris center detection. For example, the open width of eye is a good supplement for low quality videos. (3) Relative movement information is easier to be detected in inter-frame of low quality videos, which can be emphasized rather than the rough absolute position.

In almost all previous research, the basic eye-movement types such as saccades and fixations are considered the basic elements of eye movement [19], [20]. However, Heiko, Salvucci *et al.* and Goldberg *et al.* [21], [22] argued that these types are meaningless in physiology and they are not necessary for eye-movement analysis. Therefore, we suggest abandoning the use of basic eye-movement types. The feasibility of this idea has been assessed in our previous work [23]. In addition, webcam based eye movement record method becomes more natural and noninvasive compared with the method in [18], where results of electrooculography become insignificant due to vestibule-ocular reflex.

In this paper, a novel method is proposed to analyze eye movement under webcam. Different from the feature-based gaze tracking method, the feature points are used to obtain eye movement signals instead of estimating the mapping function. Compared with [23], the specific improvements can be summarized as follows. (1) The CNN is used to extract the eye feature points, where the detection results can become more robust and precise. (2) Eye movement features are also extracted by the CNN rather than artificial extracting. (3) More cues are added to analyze the eye movement patterns, which include relative displacement of iris center and variation of open width. The experimental results show that the proposed method obtains promising results. The advantages of lower invasive, lower cost and easier operation in webcam-based method are the key component for the popularization of the applications based on eye movement.

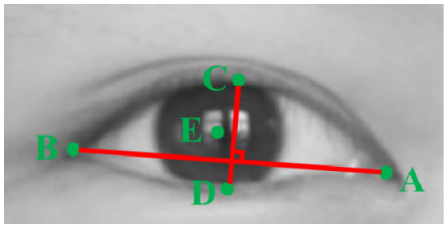

FIGURE 1. Eye feature point, A: inner corner, B: outer corner, C: the center of upper eyelid, D: the center of lower eyelid, E: the center of iris.

II. EYE FEATURE POINT DETECTION

Eye movement information can be described by eye feature point in sequential eye images. In this paper, five feature points are specifically defined and calibrated based on a unified standard. In Fig.1, the five points are shown. In particular, points A and B are inner and outer corners of the eye, respectively, and points C and D represent the centers of upper and lower eyelids, respectively. These four points are calibrated by the PB-points method introduced in our early works [24]. The location of the center of the iris (point E) is difficulty to be identified, since the video images captured by webcam are not good enough and eyelid occlusion is ubiquitous. In this paper, point E is defined as barycenter of the area surrounded by upper, lower eyelids and the visible iris outer edge.

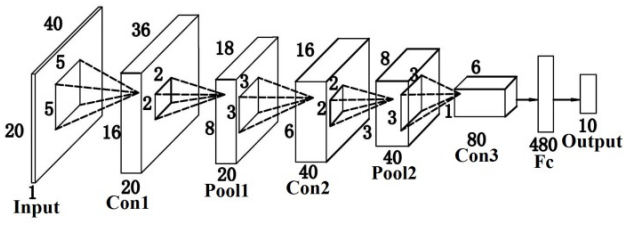

FIGURE 2. Architecture of the Points-CNN.

To detect the eye feature point, points-CNN is designed, whose architecture is shown in Figure 2. For the input image, the gray eye image is normalized to a fixed resolution

20×40. Instead of the Sigmoid and Tanh functions, the ReLU (Rectified Linear Units) activation function is used to produce the output feature of the convolutional layer with reduced computational complexity. Two max-pooling layers are used, similar to the LeNet network architecture [25].

III. EYE MOVEMENT FEATURE EXTRACTION AND EYE MOVEMENT ANALYSIS

In traditional eye movement analysis methods, two steps are necessary. Firstly, basic eye-movement types such as saccades, fixations, and blinks should be detected. Secondly, analysis modeling is established based on the detected types. The relationship between basic eye-movement types and human activity is very close. However, it is almost impossible to explicitly define a sort of basic eye-movement type [23]. In addition, these types does not make any sense in physiology [21], [22], which are not necessary. Furthermore, it is impossible to detect fast saccades (duration less than 30ms), where larger errors would be produced by basic eye-movement types detection algorithms for webcam. Therefore, it is preferable to directly use the original time-varying eye movement signals (for example the displacement of iris center) without detecting the basic eye-movement types. The feasibility of this strategy has been validated in our previous work [23].

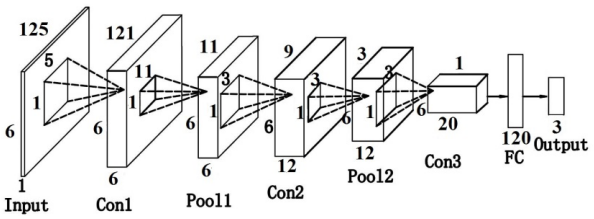

FIGURE 3. Architecture of the Behaviors-CNN.

In this paper, additional cues such as relative displacement of iris center and variation of open width have been added to analyze the eye movement patterns. These subtle cues can be quantified as feature values due to the detection of the five feature points. Moreover, artificial feature or feature combination [23] based on the original time-varying signal is no longer required. The behaviors-CNN is designed to extract more expressive eye movement features for recognizing activities from the eye movement signals, which is shown in Figure 3. The inputs include six eye movement signals, namely, vertical relative displacement (VRD) of left iris center, horizontal relative displacement (HRD) of left iris center, vertical relative displacement of right iris center, horizontal relative displacement of right iris center, variation of open width (OW) of left eye, and variation of open width of left eye. Relative displacement of iris center is a coordinate of point E relative to point A, which is insensitive to head movements. Variation of open width can be indicated by the distance between points C and D. Both the relative displacement and the open width are normalized by the eye size, which can be indicated by the distance between points A and B. Similar information of two-eyes is used to obtain improved robustness. In addition, both of the changes in vertical direction of the iris center and in open width are rather small, where the weights of these two features will be doubled before used as input to the network.

The time window of each input signal is 5 seconds. In other words, the duration of a sample is 5 second including 125 frames. A 6×125 matrix is constructed by the six types of signals as input. Unlike the input image, there is no explicit relevance among these signals. Therefore, one-dimensional convolution and one-dimensional pooling are employed in Behaviors-CNN. The ReLu activation function is used in every convolutional layer. Non-overlapping max pooling and big pooling windows (1×11) are applied to remove redundancy in slowly changing signals at the third layer.

Compared with the method in [23], additional four feature points are capable of providing more eye movement information and more precisely normalized data. Firstly, the influence of head movement can be weakened by applying relative displacement of iris center. Secondly, the variation of open width directly reflects the vertical eye movement information, since the change of open width can be more obvious than that of iris center when you look up or down. Thirdly, the change of facial size and distance to camera can be borne after normalized with the eye size.

Eye movement signals can be corrupted by feature point detection error from different sources, such as poor quality of eye image, head movement, eyelid occlusion, undeterminable outer eye corner, ambiguous center of iris and the regression model. Moreover, there are not any specific physiological meanings when saccade in a small range occurred. Therefore, an averaging filter with length of 5 is used for denoising purpose, where quality of the corrupted signals can be improved.

IV. EXPERIMENTS

In this section, the performance of the proposed feature point detection method is demonstrated, and activity recognition is used for performance evaluation of proposed eye movement analysis method. Three office-based visual activities are tested, namely, reading electronic document, watching a video and browsing the Web. These activities can only be distinguished by eye movement pattern recognition.

A. DATASETS

To the best of our knowledge, there is no available dataset with labeled eye movement for office-based visual activity recognition under webcam. For a lack of suitable data, a webcam-based eye-movement activity (WEActivity) dataset was constructed containing of almost 0.5 million frames collected from 38 subjects (two of them wearing glasses). Head shoulder sequence images were captured by two kinds of webcam (Logitech C525 and Logitech C930e) fixed on the top of the screen (daylight illumination in daytime and lamplight illumination in night). Almost one third of the data was captured by Logitech C930e (720P), where the

remainder was captured by Logitech C525 (720P). The frame rates are 25 frames per second.

1) COLLECTION PROCEDURE

Each subject is asked to perform a task while sitting in front of a screen at different time over the day. They can adjust sitting without restriction in head, body and distance. The WEActivity dataset covers a realistic variability in appearance and illumination. To avoid mislabeling, at least one supervisor is assigned to watch the whole processing during image collection, which can identify when a subject executes the allowable task.

2) STIMULATING MATERIALS

The electronic documents for reading are presented in Microsoft Office Word or PDF of Adobe. The contents of the documents cover many fields including literature, computer, news and history, where diagrams are excluded from the contents. Directindustry Web Guide or Portal web is used as the test web, where the subject can click any links. Stimulating videos mainly consists of two categories, namely, videos with and without subtitles (Chinese or English dubbing). Based on their content, there are film, teleplay, video clips, sport video and so on.

3) WEACTIVITY DATASET CHARACTERISTICS

A total of 4225-second video is selected and labeled from the whole 0.5 million frames. Some of the fruitless clips are not used. For example, some clips are not approved by the supervisor and the clips of the scene that eyes are sheltered. The approach in [23] is used to detect blink, where the frames involving blink will be removed and the video is considered to be successive before and after the blink. Taking 5 second as time window to generate data samples, we can obtain 900 video clips with non-overlapping sampling. There are 218 clips labeled as ''Reading'', 254 clips labeled as ''Browsing'', and 428 clips labeled as ''watching''. There will be over 3000 clips via overlapping sampling.

4) DATASET FOR FEATURE POINT DETECTION

A total of 2000 images are randomly selected from the WEActivity dataset (original data of 0.5 million frames) and converted to gray images. The face and eye windows of these images are detected by Haar-adaboost [26]. Moreover, 1000 left eye images and 1000 right eye images are selected and normalized to a fixed resolution of 20×40 pixels, respectively. Subsequently, manual labeling is performed for the five feature points. We selected and labeled 1754 images from the dataset as introduced in [23], which includes 144 volunteers. In view of the symmetry of binoculars, two images can be generated by flipping each labeled eye images by 180 degrees. For example, a left eye images can be transform into a right eye images simply by flipping. Therefore, there are 3754 labeled left eye images and 3754 labeled right eye images in total. In order to reduce the workload for labeling, only one network for right eye is trained, where the feature points of the left eye can be detected by rotating the eye window around the median vertical line by 180 degrees.

B. PERFORMANCE OF THE FEATURE POINT DETECTION

In particular, 3000 left eye images and 3000 right eye images are randomly selected as the training data to Points-CNN, where the remainder is used for testing. The stochastic gradient descent is used for training with an initial learning rate of 0.1. Detection precision and detection error are used for performance evaluation. Detection precision is defined by the fraction of feature points whose detection errors are below a given threshold, where the detection error is defined by

$$err = \sqrt{\frac{(x-x')^2+(y-y')2}{w.h}} \tag{1}$$

where (x, y) and (x', y') represent calibration value and detection value, respectively; w and h are the width and height of the input image, respectively. If the error of the calibration value and detection value is less than a threshold value, it can be considered to be a valid detection.

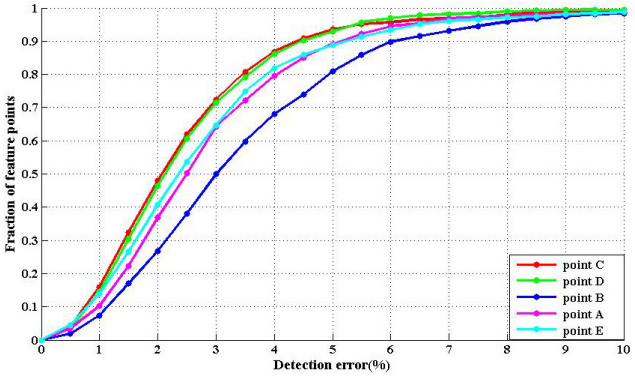

FIGURE 4. Cumulative error curves on the validation set. Points A and B are inner and outer corners of the eye respectively, points C and D respectively represent the centers of upper and lower eyelids, the center of the iris is point E.

Figure 4 shows the cumulative error curves of the proposed method. When the detection error is 5% (about 2 pixels), the detection accuracies of points B, E, A, D and C are 80.9%, 88.7%, 89.7%, 92.8%, 93.5%, respectively. When the detection error is 10% (about 4 pixels), the detection accuracies of points B, E, A, D and C are 98.3%, 98.7%, 98.7%, 99.5%, 99.4%, respectively. The precisions of different feature point detection are unbalanced and the difficulties of detection are inconsistent. The effects can become more significant especially when the detection error is from 2% to 6%. Compared with the other feature points, the performance on the outer eye corner is worse, since the characteristic of the point is rather inconspicuous and is difficult to be ascertained by human. This feature point detection method will be used in the WEActivity dataset.

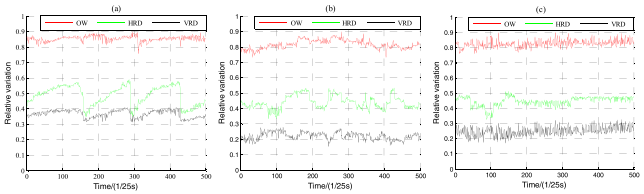

FIGURE 5. Comparison of different eye movement signals of one subject's right eye. OW refers to the variation of the open width, HRD refers to the horizontal relative displacement, and VRD refers to the vertical relative displacement.

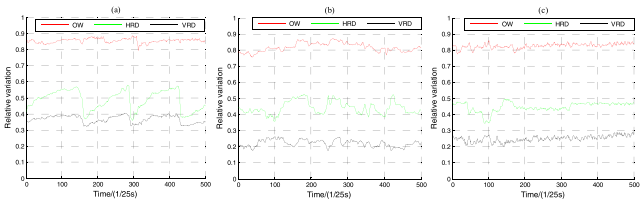

FIGURE 6. Corresponding signals after filtering. OW refers to the variation of the open width, HRD refers to the horizontal relative displacement, and VRD refers to the vertical relative displacement.

C. ACTIVITY RECOGNITION BASED ON THE PROPOSED EYE MOVEMENT ANALYSIS METHOD

Since the accuracy of the eye movement signals is bounded by the quality of the webcam and limitation of image capture scene, gaze estimation under webcam is rather coarse for gaze based human-computer interactions. Therefore, eye movement based visual tasks is proposed for evaluating the performance of the presented eye movement analysis method.

In particular, six types of eye movement signals can be constructed by five eye feature points, namely, vertical relative displacement of left iris center, horizontal relative displacement of left iris center, vertical relative displacement of right iris center, horizontal relative displacement of right iris center, variation of the open width of left eye, and variation of the open width of left eye. In Figure 5, parts of the signals are shown for different activities of one subject. As shown in Figure 6, the outlines of the corresponding signals become more distinct, where non-informative details can be removed after filtering. It can be clearly observed that the relative displacement signals of reading have clear characteristics of sawtooth like wave. The other two types of signals seem to be both intricate and ruleless. However, there are more fixations when watching a video compared to browsing the Web. To some extent, these characteristics can indicate that different activities can be separated based on eye movement signals.

In this experiment, 700 video clips are randomly chosen to train the Behaviors-CNN, which are sampled from the 900 video clips with non-overlapping sampling strategy. The remaining 200 clips are used for testing. The stochastic gradient descent is used for training with an initial learning rate of 0.1. Convergence can be obtained with around 200 iterations.

In Table 1, the mean recognition results of 10 repeated experiments are shown. It can be found that ''Reading'' can be recognized with a high precision of 87.5 percent and recall of 89.4 percent due to the distinctiveness of the feature. There are often more fixations during watching a video, where the characteristics are less obvious than that of ''Reading'', but more obvious than that of browsing the Web. Therefore, browsing can be recognized worse with a precision of 72.9 percent and a recall of 71.7 percent. In addition, reading activity will be involved when one watches a video (particularly the video with subtitle) or browses the Web. Therefore, it is incidental that these two activities will be mistakenly recognized as reading.

TABLE 1. Recognition precision and recall of different activities.

Rank	Precision	Recall
Reading	87.5%	89.4%
Browsing the Web	72.9%	71.7%
Watching a video	83.5%	83.5%
Mean	81.3%	81.5%

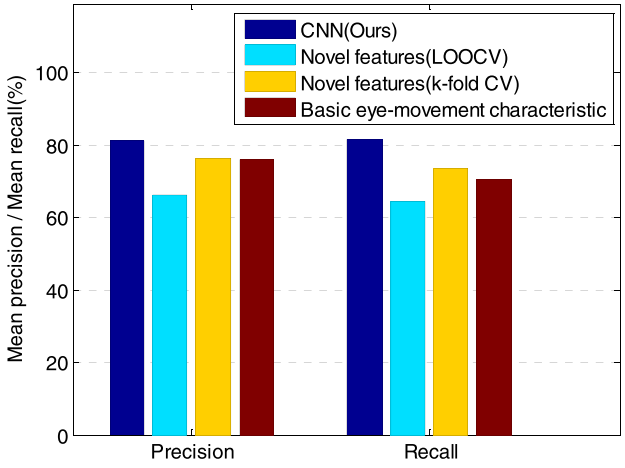

FIGURE 7. Comparison of the different models. CNN and five feature points are used in our method without any artificial feature extraction. Novel features represent the model presented by [23], where LOOCV is ''leave one out cross validation'' and k-fold CV is ''k-fold cross-validation''. Basic eye-movement characteristics represent the model in [18]. The bars indicate the corresponding mean precision or mean recall.

In Figure 7, different models of eye movement based activities recognition are compared. As can be seen from this figure, the performance of the proposed method is very promising. This is because both the feature points and eye movement features are extracted by using CNN to avoid the error from either artificial feature extraction or detection of the basic eye-movement characteristic. In addition, compared with the method in [23], more significant feature points have been used. Compared with the method in [2], webcam based mode is far more friendly, natural and convenient than electrooculography. It should be noted that the results of our model and the model in [23] are all tested on the WEActivity dataset. In particular, the results of [18] are cited because their data is collected by electrooculogram. The novel

TABLE 2. Recognition results of reading and non-reading.

Test number	Reading		Non-reading	
	Precision	Recall	Precision	Recall
1	89.09%	92.45%	97.24%	95.92%
2	91.49%	95.56%	98.69%	97.42%
3	90.74%	94.23%	97.95%	96.62%
4	92.59%	89.29%	95.89%	97.22%
5	83.67%	89.13%	97.98%	94.81%
6	81.13%	91.49%	97.28%	93.46%
7	95.74%	97.83%	99.35%	98.70%
8	93.75%	97.83%	99.34%	98.05%
9	81.13%	91.49%	97.28%	93.46%
10	91.49%	91.49%	97.39%	97.39%
Mean	88.95%	92.4%	97.53%	96.30%

features (leave one out cross validation, LOOCV) model performs poorer than the novel features (k-fold cross-validation, k-fold CV) due to calibration-free characteristic. Although eye movement signals recorded by electrooculogram can be more precise than that obtained by webcam, the results of the model presented in [18] involve five kinds of activities, where diagrams are included in its stimulus material of reading. In our experiments, reading only represents reading the text, while watching a diagram is classified into watching video.

Since reading is ubiquitous in real-world and office-based activities, an experiment is designed to distinguish ''reading'' and ''non-reading'' tasks for evaluating the eye movement analysis method. Similarly, 700 video clips are randomly selected from the 900 video clips to train the Behaviors-CNN (including 2 outputs). The remaining 200 clips are used for testing. Training is performed by stochastic gradient descent with an initial learning rate of 0.1. In Table 2, the recognition results of 10 repeated experiments are presented. The mean value of precise and recall of the two types are around 93 percent and 94 percent, respectively. The precision for reading is the lowest, since there is reading behavior in non-reading data such as reading texts in the Web and subtitles in videos. If reading during browsing and watching video can be correctly labeled, the performance will become much better. In most cases, reading cannot be recognized as good as non-reading due to the imbalanced training data. In particular, there are often more training data for non-reading cases in most experiments. By adjusting the training process or proportion of different sorts based on different applied scenarios, it can be guaranteed to have an indicator of over 95 percent. These results have validated that the proposed model is competent for the applications related to reading detection.

V. CONCLUSION

In this paper, a novel eye movement analysis model is proposed. To verify the feasibility, a webcam-based visual activity dataset is collected and constructed. The proposed CNN-based model can outperform other state-of-the-art methods in three office activities recognition tasks. Although the accuracy of eye tracking is limited by the quality of webcam, it is revealed that the proposed webcam-based eye movement analysis method can be successfully employed to recognize human visual activities without gaze estimation and detection of the basic eye-movement types. Moreover, it can be shown that accurate recognition results can be achieved by using CNN, which is capable of extracting more representative internal feature from original time-varying eye movement signals. In particular, this model provides a new way for eye movement analysis by detecting the feature point and classifying the original time-varying eye movement signals using CNN.

REFERENCES

[1] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*. London, U.K.: Springer, 2007, pp. 51–58.
[2] K. Rayner, T. J. Smith, G. L. Malcolm, and J. M. Henderson, ''Eye movements and visual encoding during scene perception,'' *Psychol. Sci.*, vol. 20, no. 1, pp. 6–10, 2009.
[3] D. Schneider, A. P. Bayliss, S. I. Becker, and P. E. Dux, ''Eye movements reveal sustained implicit processing of others' mental states,'' *J. Experim. Psychol., Gen.*, vol. 141, no. 3, p. 433, 2012.
[4] H. Jarodzka *et al.*, ''Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases,'' in *Proc. 32nd Annu. Meet. Cognit. Sci. Soc.*, 2010, pp. 1703–1708.
[5] C. Tonkin, A. D. Ouzts, and A. T. Duchowski, ''Eye tracking within the packaging design workflow: Interaction with physical and virtual shelves,'' in *Proc. 1st Conf. Novel Gaze-Controlled Appl.*, New York, NY, USA, 2011. Art. no. 3.
[6] S. V. Sheela and P. A. Vijaya, ''Iris recognition methods-survey,'' *Int. J. Comput. Appl.*, vol. 3, no. 5, pp. 19–25, 2010.
[7] R. Biedert, G. Buscher, and A. Dengel, ''The eyeBook—Using eye tracking to enhance the reading experience,'' *Inform.-Spektrum*, vol. 33, no. 3, pp. 272–281, 2010.
[8] D. W. Hansen and Q. Ji, ''In the eye of the beholder: A survey of models for eyes and gaze,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.
[9] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, ''Appearance-based gaze estimation in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4511–4520.
[10] D. W. Hansen, J. P. Hansen, M. Nielsen, A. S. Johansen, M. B. Stegmann, ''Eye typing using Markov and active appearance models,'' in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Dec. 2002, pp. 132–136.
[11] K.-N. Kim and R. S. Ramakrishna, ''Vision-based eye-gaze tracking for human computer interface,'' in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, vol. 2. Oct. 1999, pp. 324–329.
[12] J.-G. Wang, E. Sung, and R. Venkateswarlu, ''Eye gaze estimation from a single image of one eye,'' in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 136–143.
[13] W. Zhang, T.-N. Zhang, and S.-J. Chang, ''Eye gaze estimation from the elliptical features of one iris,'' *Opt. Eng.*, vol. 50, no. 4, pp. 047003-1–047003-9, 2011.
[14] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, ''Adaptive linear regression for appearance-based gaze estimation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.
[15] K.-H. Tan, D. J. Kriegman, and N. Ahuja, ''Appearance-based eye gaze estimation,'' in *Proc. 6th IEEE Workshop Appl. Comput. Vis. (WACV)*, Dec. 2002, pp. 191–195.
[16] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[17] K. Krafka *et al.*, ''Eye tracking for everyone,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2176–2184.
[18] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, ''Eye movement analysis for activity recognition using electrooculography,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 741–753, Apr. 2011.

[19] B. R. Manor and E. Gordon, "Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks," *J. Neurosci. Methods*, vol. 128, nos. 1–2, pp. 85–93, 2003.
[20] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, 2002.
[21] H. Drewes, "Eye gaze tracking for human computer interaction," Ph.D. dissertation, Dept. Comput. Sci. Stat., Ludwig Maximilian Univ. Munich, Munich, Germany, 2010.
[22] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, 2000, pp. 71–78.
[23] C.-N. Meng, J.-J. Bai, T.-N. Zhang, R.-B. Liu, and S.-J. Chang, "Eye movement analysis for activity recognition based on one Web camera," *Acta Phys. Sinica*, vol. 62, no. 17, pp. 174203-1–174203-8, 2013.
[24] C. Meng *et al.*, "Eye feature points detection by CNN with strict geometric constraint," *Proc. SPIE*, vol. 10033, pp. 1003340-1–1003340-5, Aug. 2016.
[25] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[26] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

CHUNNING MENG received the B.S. and M.S. degrees in physical electronics from Yantai University in 2007 and 2010, the Ph.D. degrees in optical engineering from Nankai University in 2013. He is currently an Associate Professor of computer science and technology, China Maritime Police Academy. He has authored or co-authored over 30 technical papers in the field of digital image processing, pattern recognition, and neural networks.

XUEPENG ZHAO received the B.S. degree in optic information science and technology from Hebei University, Baoding, China, in 2014, and the M.S. degree in optical engineering from Nankai University, Tianji, China, in 2017. From 2014 to 2017, he studied with the State Key Laboratory of Optical Information, Nankai University, Tianji, China. His research interest included image processing, pattern recognition, and deep learning.