21^{st} *Computer Vision Winter Workshop*
Luka Čehovin, Rok Mandeljc, Vitomir Štruc (eds.)
Rimske Toplice, Slovenia, February 3–5, 2016

Real-Time Eye Blink Detection using Facial Landmarks

Tereza Soukupová and Jan Čech
Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague
{soukuter,cechj}@cmp.felk.cvut.cz

Abstract. *A real-time algorithm to detect eye blinks in a video sequence from a standard camera is proposed. Recent landmark detectors, trained on in-the-wild datasets exhibit excellent robustness against a head orientation with respect to a camera, varying illumination and facial expressions. We show that the landmarks are detected precisely enough to reliably estimate the level of the eye opening. The proposed algorithm therefore estimates the landmark positions, extracts a single scalar quantity – eye aspect ratio (EAR) – characterizing the eye opening in each frame. Finally, an SVM classifier detects eye blinks as a pattern of EAR values in a short temporal window. The simple algorithm outperforms the state-of-the-art results on two standard datasets.*

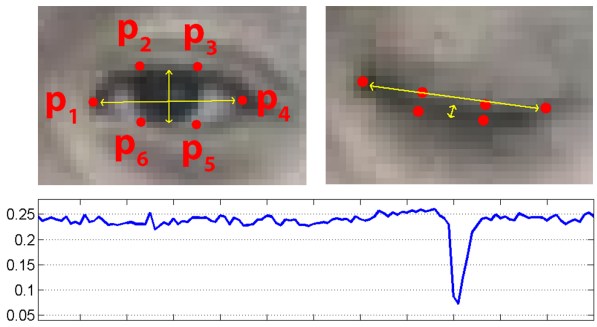


Figure 1: Open and closed eyes with landmarks p_i automatically detected by [1]. The eye aspect ratio EAR in Eq. (1) plotted for several frames of a video sequence. A single blink is present.

1. Introduction

Detecting eye blinks is important for instance in systems that monitor a human operator vigilance, e.g. driver drowsiness [5, 13], in systems that warn a computer user staring at the screen without blinking for a long time to prevent the dry eye and the computer vision syndromes [17, 7, 8], in human-computer interfaces that ease communication for disabled people [15], or for anti-spoofing protection in face recognition systems [11].

Existing methods are either active or passive. Active methods are reliable but use special hardware, often expensive and intrusive, e.g. infrared cameras and illuminators [2], wearable devices, glasses with a special close-up cameras observing the eyes [10]. While the passive systems rely on a standard remote camera only.

Many methods have been proposed to automatically detect eye blinks in a video sequence. Several methods are based on a *motion estimation* in the eye region. Typically, the face and eyes are detected by a Viola-Jones type detector. Next, motion in the eye area is estimated from optical flow, by sparse tracking [7, 8], or by frame-to-frame intensity differencing and adaptive thresholding. Finally, a decision is made whether the eyes are or are not covered by eyelids [9, 15]. A different approach is to infer the *state of the eye opening from a single image*, as e.g. by correlation matching with open and closed eye templates [4], a heuristic horizontal or vertical image intensity projection over the eye region [5, 6], a parametric model fitting to find the eyelids [18], or active shape models [14].

A major drawback of the previous approaches is that they usually implicitly impose too strong requirements on the setup, in the sense of a relative face-camera pose (head orientation), image resolution, illumination, motion dynamics, etc. Especially the heuristic methods that use raw image intensity are likely to be very sensitive despite their real-time performance.

However nowadays, robust real-time *facial landmark detectors* that capture most of the characteristic points on a human face image, including eye corners and eyelids, are available, see Fig. 1. Most of the state-of-the-art landmark detectors formulate a regression problem, where a mapping from an image into landmark positions [16] or into other landmark parametrization [1] is learned. These modern landmark detectors are trained on "in-the-wild datasets" and they are thus robust to varying illumination, various facial expressions, and moderate non-frontal head rotations. An average error of the landmark localization of a state-of-the-art detector is usually below five percent of the inter-ocular distance. Recent methods run even significantly super real-time [12].

Therefore, we propose a simple but efficient algorithm to detect eye blinks by using a recent facial landmark detector. A single scalar quantity that reflects a level of the eye opening is derived from the landmarks. Finally, having a per-frame sequence of the eye opening estimates, the eye blinks are found by an SVM classifier that is trained on examples of blinking and non-blinking patterns.

Facial segmentation model presented in [14] is similar to the proposed method. However, their system is based on active shape models with reported processing time of about 5 seconds per frame for the segmentation, and the eye opening signal is normalized by statistics estimated by observing a longer sequence. The system is thus usable for offline processing only. The proposed algorithm runs real-time, since the extra costs of the eye opening from landmarks and the linear SVM are negligible.

The contributions of the paper are:

1. Ability of two state-of-the-art landmark detectors [1, 16] to reliably distinguish between the open and closed eye states is quantitatively demonstrated on a challenging in-the-wild dataset and for various face image resolutions.

2. A novel real-time eye blink detection algorithm which integrates a landmark detector and a classifier is proposed. The evaluation is done on two standard datasets [11, 8] achieving state-of-the-art results.

The rest of the paper is structured as follows: The algorithm is detailed in Sec. 2, experimental validation and evaluation is presented in Sec. 3. Finally, Sec. 4 concludes the paper.

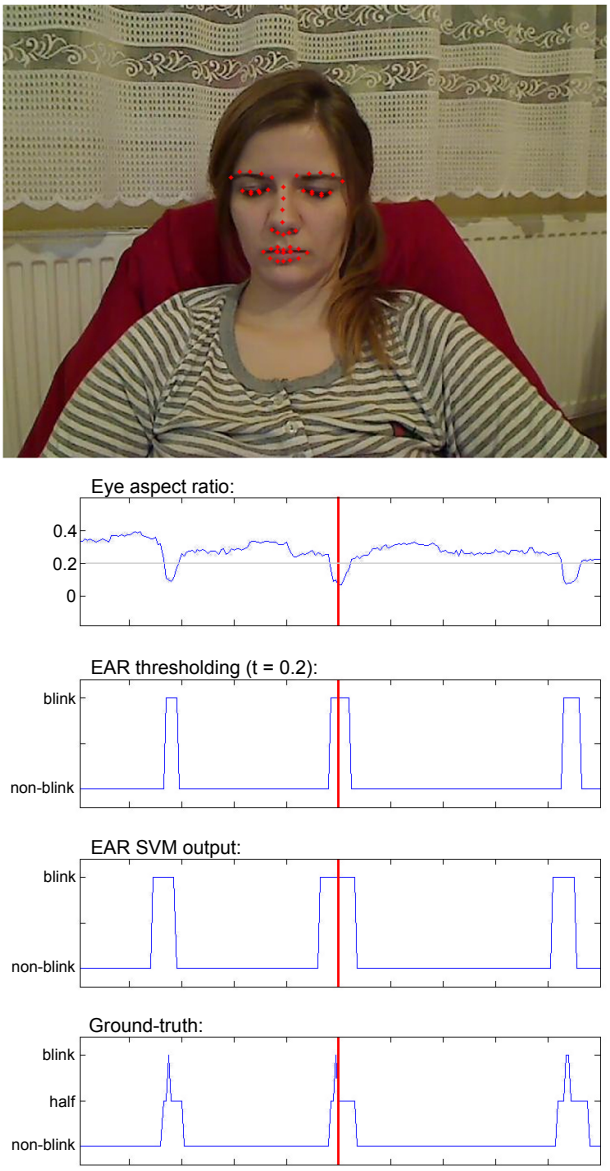


Figure 2: Example of detected blinks. The plots of the eye aspect ratio EAR in Eq. (1), results of the EAR thresholding (threshold set to 0.2), the blinks detected by EAR SVM and the ground-truth labels over the video sequence. Input image with detected landmarks (depicted frame is marked by a red line).

2. Proposed method

The eye blink is a fast closing and reopening of a human eye. Each individual has a little bit different pattern of blinks. The pattern differs in the speed of closing and opening, a degree of squeezing the eye and in a blink duration. The eye blink lasts approxi-

mately 100-400 ms.

We propose to exploit state-of-the-art facial landmark detectors to localize the eyes and eyelid contours. From the landmarks detected in the image, we derive the eye aspect ratio (EAR) that is used as an estimate of the eye opening state. Since the per-frame EAR may not necessarily recognize the eye blinks correctly, a classifier that takes a larger temporal window of a frame into account is trained.

2.1. Description of features

For every video frame, the eye landmarks are detected. The eye aspect ratio (EAR) between height and width of the eye is computed.

$$\text{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}, \qquad (1)$$

where $p_1, \ldots, p_6$ are the 2D landmark locations, depicted in Fig. 1.

The EAR is mostly constant when an eye is open and is getting close to zero while closing an eye. It is partially person and head pose insensitive. Aspect ratio of the open eye has a small variance among individuals and it is fully invariant to a uniform scaling of the image and in-plane rotation of the face. Since eye blinking is performed by both eyes synchronously, the EAR of both eyes is averaged. An example of an EAR signal over the video sequence is shown in Fig. 1, 2, 7.

A similar feature to measure the eye opening was suggested in [9], but it was derived from the eye segmentation in a binary image.

2.2. Classification

It generally does not hold that low value of the EAR means that a person is blinking. A low value of the EAR may occur when a subject closes his/her eyes intentionally for a longer time or performs a facial expression, yawning, etc., or the EAR captures a short random fluctuation of the landmarks.

Therefore, we propose a classifier that takes a larger temporal window of a frame as an input. For the 30fps videos, we experimentally found that ± 6 frames can have a significant impact on a blink detection for a frame where an eye is the most closed when blinking. Thus, for each frame, a 13-dimensional feature is gathered by concatenating the EARs of its ± 6 neighboring frames.

This is implemented by a linear SVM classifier (called EAR SVM) trained from manually annotated sequences. Positive examples are collected as ground-truth blinks, while the negatives are those that are sampled from parts of the videos where no blink occurs, with 5 frames spacing and 7 frames margin from the ground-truth blinks. While testing, a classifier is executed in a scanning-window fashion. A 13-dimensional feature is computed and classified by EAR SVM for each frame except the beginning and ending of a video sequence.

3. Experiments

Two types of experiments were carried out: The experiments that measure accuracy of the landmark detectors, see Sec. 3.1, and the experiments that evaluate performance of the whole eye blink detection algorithm, see Sec 3.2.

3.1. Accuracy of landmark detectors

To evaluate accuracy of tested landmark detectors, we used the 300-VW dataset [19]. It is a dataset containing 50 videos where each frame has associated a precise annotation of facial landmarks. The videos are "in-the-wild", mostly recorded from a TV.

The purpose of the following tests is to demonstrate that recent landmark detectors are particularly robust and precise in detecting eyes, i.e. the eye-corners and contour of the eyelids. Therefore we prepared a dataset, a subset of the 300-VW, containing sample images with both open and closed eyes. More precisely, having the ground-truth landmark annotation, we sorted the frames for each subject by the eye aspect ratio (EAR in Eq. (1)) and took 10 frames of the highest ratio (eyes wide open), 10 frames of the lowest ratio (mostly eyes tightly shut) and 10 frames sampled randomly. Thus we collected 1500 images. Moreover, all the images were later subsampled (successively 10 times by factor 0.75) in order to evaluate accuracy of tested detectors on small face images.

Two state-of-the-art landmark detectors were tested: Chehra [1] and Intraface [16]. Both run in real-time[1]. Samples from the dataset are shown in Fig. 3. Notice that faces are not always frontal to the camera, the expression is not always neutral, people are often emotionally speaking or smiling, etc. Sometimes people wear glasses, hair may occasionally partially occlude one of the eyes. Both detectors perform generally well, but the Intraface is more robust to very small face images, sometimes at impressive extent as shown in Fig. 3.

[1]Intraface runs in 50 Hz on a standard laptop.

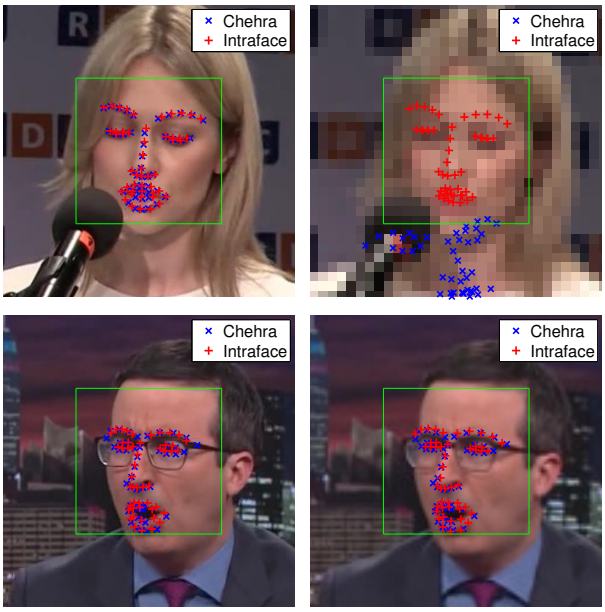


Figure 3: Example images from the 300-VW dataset with landmarks obtained by Chehra [1] and Intraface [16]. Original images (left) with inter-ocular distance (IOD) equal to 63 (top) and 53 (bottom) pixels. Images subsampled (right) to IOD equal to 6.3 (top) and 17 (bottom).

Quantitatively, the accuracy of the landmark detection for a face image is measured by the average relative landmark localization error, defined as usually

$$\epsilon = \frac{100}{\kappa N} \sum_{i=1}^{N} ||x_i - \hat{x}_i||_2, \qquad (2)$$

where x_i is the ground-truth location of landmark i in the image, $\hat{x}_i$ is an estimated landmark location by a detector, N is a number of landmarks and normalization factor κ is the inter-ocular distance (IOD), i.e. Euclidean distance between eye centers in the image.

First, a standard cumulative histogram of the average relative landmark localization error ϵ was calculated, see Fig. 4, for a complete set of 49 landmarks and also for a subset of 12 landmarks of the eyes only, since these landmarks are used in the proposed eye blink detector. The results are calculated for all the original images that have average IOD around 80 px, and also for all "small" face images (including subsampled ones) having IOD $\leq$ 50 px. For all landmarks, Chehra has more occurrences of very small errors (up to 5 percent of the IOD), but Intraface is more robust having more occurrences of errors below 10 percent of the IOD. For eye landmarks only,

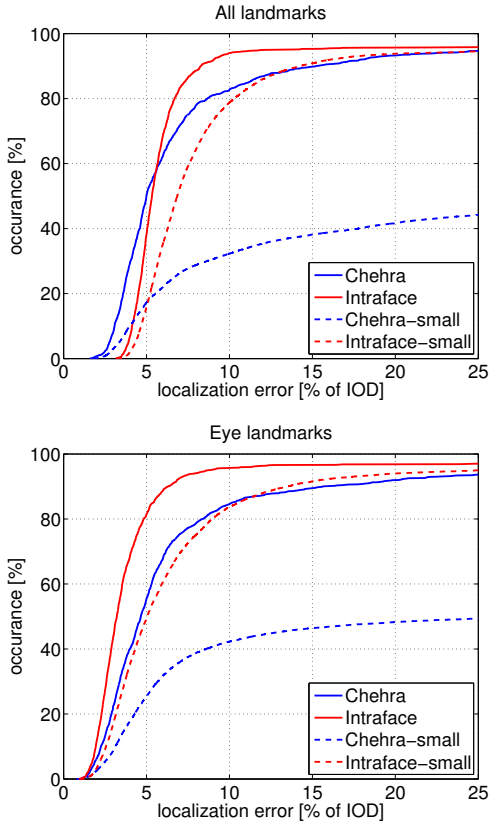


Figure 4: Cumulative histogram of average localization error of all 49 landmarks (top) and 12 landmarks of the eyes (bottom). The histograms are computed for original resolution images (solid lines) and a subset of small images (IOD $\leq$ 50 px).

the Intraface is always more precise than Chehra. As already mentioned, the Intraface is much more robust to small images than Chehra. This behaviour is further observed in the following experiment.

Taking a set of all 15k images, we measured a mean localization error μ as a function of a face image resolution determined by the IOD. More precisely, $\mu = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \epsilon_j$, i.e. average error over set of face images $\mathcal{S}$ having the IOD in a given range. Results are shown in Fig. 5. Plots have errorbars of standard deviation. It is seen that Chehra fails quickly for images with IOD $<$ 20 px. For larger faces, the mean error is comparable, although slightly better for Intraface for the eye landmarks.

The last test is directly related to the eye blink detector. We measured accuracy of EAR as a function of the IOD. Mean EAR error is defined as a mean absolute difference between the true and the estimated EAR. The plots are computed for two subsets: closed/closing (average true ratio 0.05 ± 0.05)

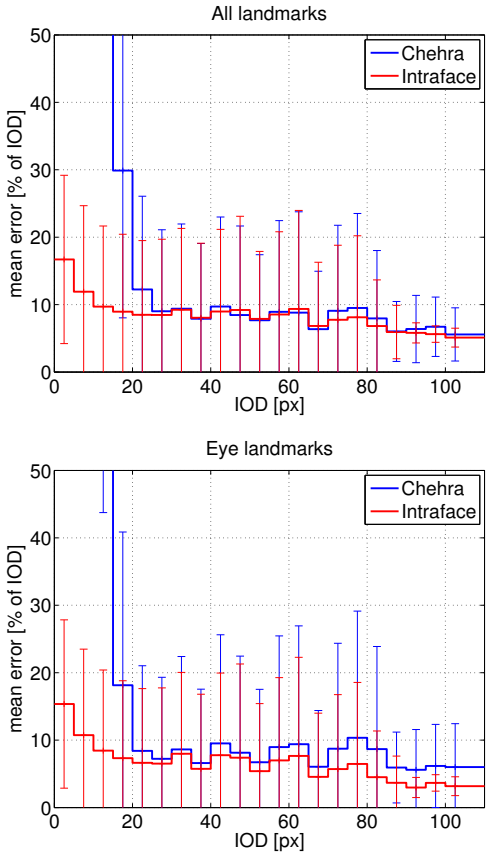


Figure 5: Landmark localization accuracy as a function of the face image resolution computed for all landmarks and eye landmarks only.

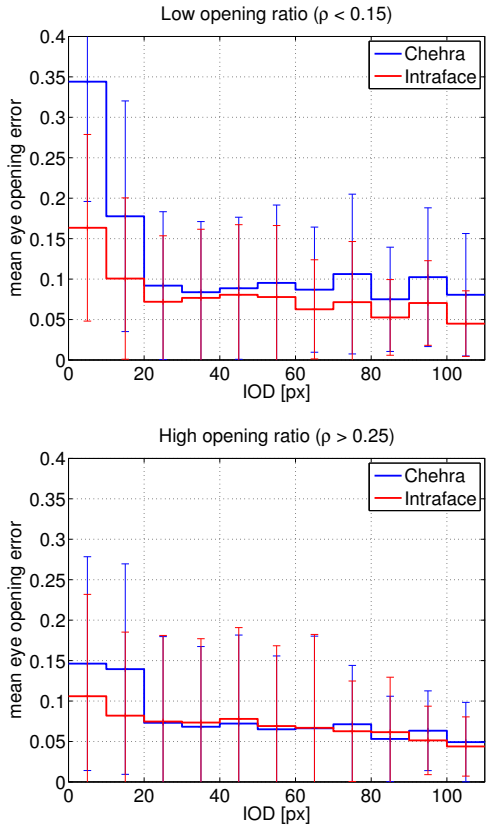


Figure 6: Accuracy of the eye-opening ratio as a function of the face image resolution. Top: for images with small true ratio (mostly closing/closed eyes), and bottom: images with higher ratio (open eyes).

and open eyes (average true ratio 0.4 ± 0.1). The error is higher for closed eyes. The reason is probably that both detectors are more likely to output open eyes in case of a failure. It is seen that ratio error for IOD $<$ 20 px causes a major confusion between open/close eye states for Chehra, nevertheless for larger faces the ratio is estimated precisely enough to ensure a reliable eye blink detection.

3.2. Eye blink detector evaluation

We evaluate on two standard databases with ground-truth annotations of blinks. The first one is ZJU [11] consisting of 80 short videos of 20 subjects. Each subject has 4 videos: 2 with and 2 without glasses, 3 videos are frontal and 1 is an upward view. The 30fps videos are of size 320 $\times$ 240 px. An average video length is 136 frames and contains about 3.6 blinks in average. An average IOD is 57.4 pixels. In this database, subjects do not perform any noticeable facial expressions. They look straight into the camera at close distance, almost do not move, do not either smile nor speak. A ground-truth blink is defined by its beginning frame, peak frame and ending frame. The second database Eyeblink8 [8] is more challenging. It consists of 8 long videos of 4 subjects that are smiling, rotating head naturally, covering face with hands, yawning, drinking and looking down probably on a keyboard. These videos have length from 5k to 11k frames, also 30fps, with a resolution 640 $\times$ 480 pixels and an average IOD 62.9 pixels. They contain about 50 blinks on average per video. Each frame belonging to a blink is annotated by half-open or close state of the eyes. We consider half blinks, which do not achieve the close state, as full blinks to be consistent with the ZJU.

Besides testing the proposed EAR SVM methods, that are trained to detect the specific blink pattern, we compare with a simple baseline method, which only thresholds the EAR in Eq. (1) values. The EAR SVM classifiers are tested with both landmark detectors Chehra [1] and Intraface [16].

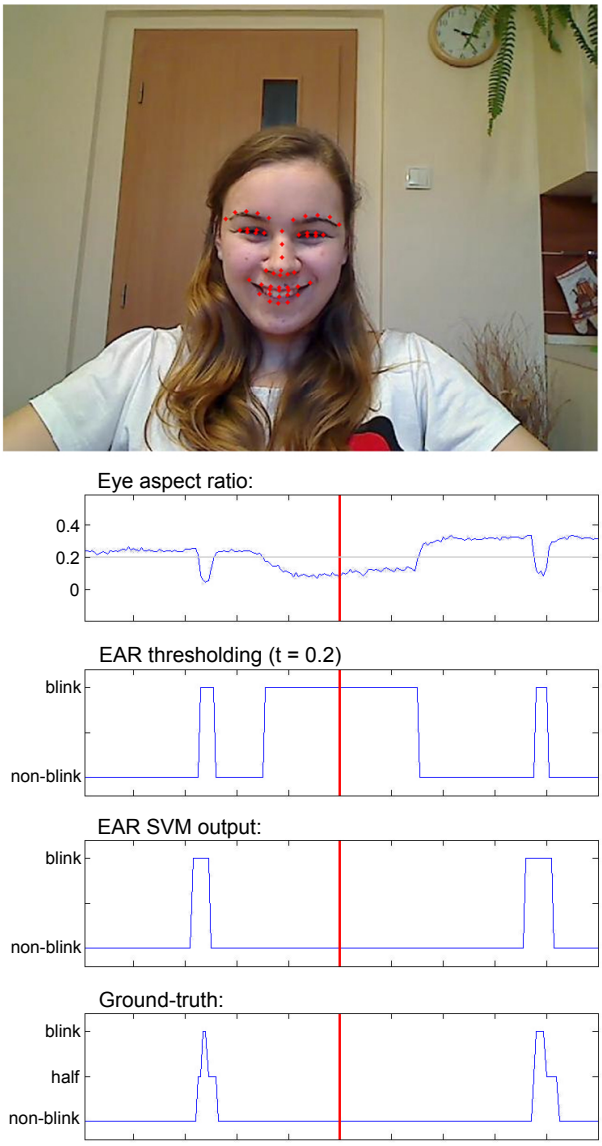


Figure 7: Example of detected blinks where the EAR thresholding fails while EAR SVM succeeds. The plots of the eye aspect ratio EAR in Eq. (1), results of the EAR thresholding (threshold set to 0.2), the blinks detected by EAR SVM and the ground-truth labels over the video sequence. Input image with detected landmarks (depicted frame is marked by a red line).

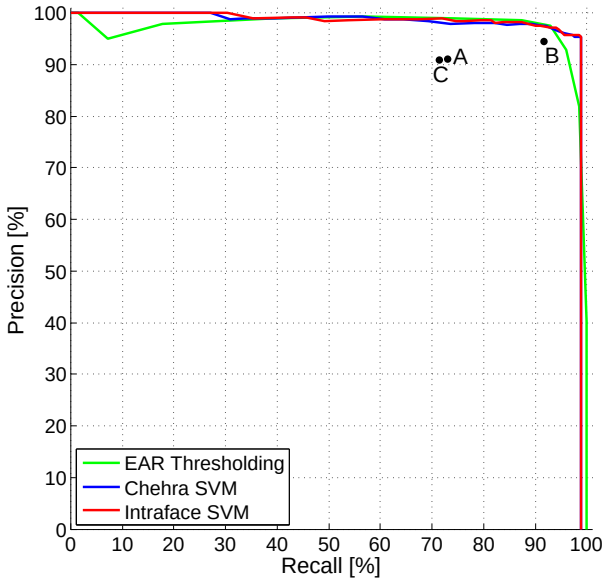
The experiment with EAR SVM is done in a cross-dataset fashion. It means that the SVM classifier is trained on the Eyeblink8 and tested on the ZJU and vice versa.

To evaluate detector accuracy, predicted blinks are compared with the ground-truth blinks. The number of true positives is determined as a number of the ground-truth blinks which have a non-empty intersection with detected blinks. The number of false negatives is counted as a number of the ground-truth blinks which do not intersect detected blinks. The number of false positives is equal to the number of detected blinks minus the number of true positives plus a penalty for detecting too long blinks. The penalty is counted only for detecting blinks twice longer then an average blink of length A. Every long blink of length L is counted $\frac{L}{A}$ times as a false positive. The number of all possibly detectable blinks is computed as number of frames of a video sequence divided by subject average blink length following Drutarovsky and Fogelton [8].

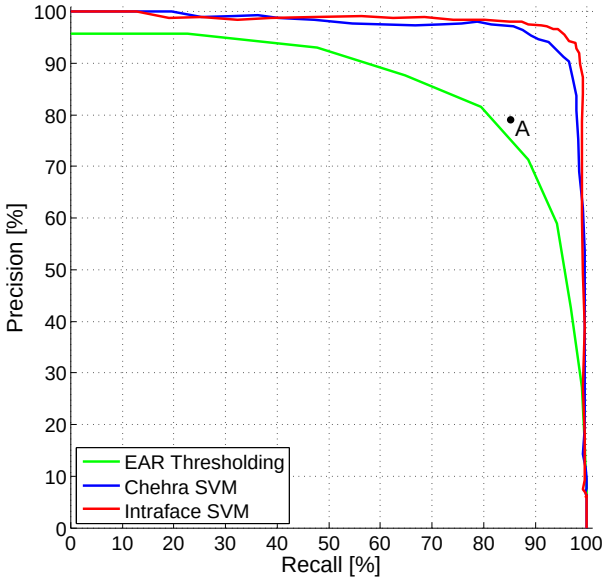
The ZJU database appears relatively easy. It mostly holds that every eye closing is an eye blink. Consequently, the precision-recall curves shown in Fig. 8a of the EAR thresholding and both EAR SVM classifiers are almost identical. These curves were calculated by spanning a threshold of the EAR and SVM output score respectively. All our methods outperform other detectors [9, 8, 5]. The published methods presented the precision and the recall for a single operation point only, not the precision-recall curve. See Fig. 8a for comparison.

The precision-recall curves in Fig. 8b shows evaluation on the Eyeblink8 database. We observe that in this challenging database the EAR thresholding lags behind both EAR SVM classifiers. The thresholding fails when a subject smiles (has narrowed eyes - see an example in Fig. 7), has a side view or when the subject closes his/her eyes for a time longer than a blink duration. Both SVM detectors performs much better, the Intraface detector based SVM is even a little better than the Chehra SVM. Both EAR SVM detectors outperform the method by Drutarovsky and Fogelton [8] by a significant margin.

Finally, we measured a dependence of the whole blink detector accuracy on the average IOD over the dataset. Every frame of the ZJU database was subsampled to 90%, 80%, ..., 10% of its original resolution. Both Chehra-SVM and Intraface-SVM were used for evaluation. For each resolution, the area under the precision-recall curve (AUC) was computed. The result is shown in Fig. 9. We can see that with Chehra landmarks the accuracy remains very high until average IOD is about 30 px. The detector fails on images with the IOD $<$ 20 px. Intraface landmarks are much better in low resolutions. This confirms our previous study on the accuracy of landmarks in Sec. 3.1.



(a) ZJU



(b) Eyeblink8

Figure 8: Precision-recall curves of the EAR thresholding and EAR SVM classifiers measured on (a) the ZJU and (b) the Eyeblink8 databases. Published results of methods A - Drutarovsky and Fogelton [8], B - Lee et al. [9], C - Danisman et al. [5] are depicted.

4. Conclusion

A real-time eye blink detection algorithm was presented. We quantitatively demonstrated that regression-based facial landmark detectors are precise enough to reliably estimate a level of eye openness. While they are robust to low image quality (low image resolution in a large extent) and in-the-wild

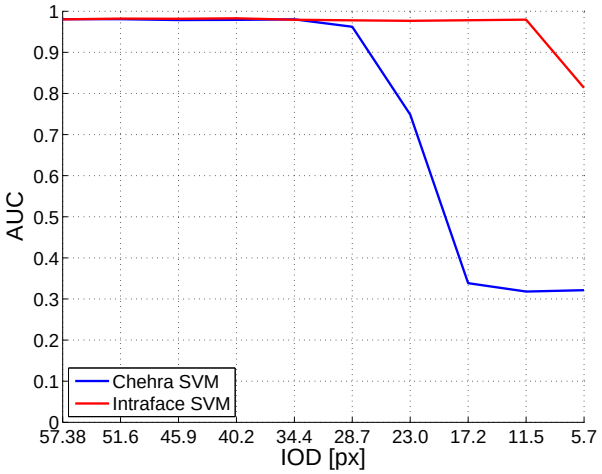


Figure 9: Accuracy of the eye blink detector (measured by AUC) as a function of the image resolution (average IOD) when subsampling the ZJU dataset.

phenomena as non-frontality, bad illumination, facial expressions, etc.

State-of-the-art on two standard datasets was achieved using the robust landmark detector followed by a simple eye blink detection based on the SVM. The algorithm runs in real-time, since the additional computational costs for the eye blink detection are negligible besides the real-time landmark detectors.

The proposed SVM method that uses a temporal window of the eye aspect ratio (EAR), outperforms the EAR thresholding. On the other hand, the thresholding is usable as a single image classifier to detect the eye state, in case that a longer sequence is not available.

We see a limitation that a fixed blink duration for all subjects was assumed, although everyone's blink lasts differently. The results could be improved by an adaptive approach. Another limitation is in the eye opening estimate. While EAR is estimated from a 2D image, it is fairly insensitive to a head orientation, but may lose discriminability for out of plane rotations. A solution might be to define the EAR in 3D. There are landmark detectors that estimate a 3D pose (position and orientation) of a 3D model of landmarks, e.g. [1, 3].

Acknowledgment

The research was supported by CTU student grant SGS15/155/OHK3/2T/13.

References

[1] A. Asthana, S. Zafeoriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2, 3, 4, 5, 7

[2] L. M. Bergasa, J. Nuevo, M. A. Sotelo, and M. Vazquez. Real-time system for monitoring driver vigilance. In *IEEE Intelligent Vehicles Symposium*, 2004. 1

[3] J. Cech, V. Franc, and J. Matas. A 3D approach to facial landmarks: Detection, refinement, and tracking. In *Proc. International Conference on Pattern Recognition*, 2014. 7

[4] M. Chau and M. Betke. Real time eye tracking and blink detection with USB cameras. Technical Report 2005-12, Boston University Computer Science, May 2005. 1

[5] T. Danisman, I. Bilasco, C. Djeraba, and N. Ihaddadene. Drowsy driver detection system using eye blink patterns. In *Machine and Web Intelligence (ICMWI)*, Oct 2010. 1, 6, 7

[6] H. Dinh, E. Jovanov, and R. Adhami. Eye blink detection using intensity vertical projection. In *International Multi-Conference on Engineering and Technological Innovation, IMETI 2012*. 1

[7] M. Divjak and H. Bischof. Eye blink based fatigue detection for prevention of computer vision syndrome. In *IAPR Conference on Machine Vision Applications*, 2009. 1

[8] T. Drutarovsky and A. Fogelton. Eye blink detection using variance of motion vectors. In *Computer Vision - ECCV Workshops*. 2014. 1, 2, 5, 6, 7

[9] W. H. Lee, E. C. Lee, and K. E. Park. Blink detection robust to various facial poses. *Journal of Neuroscience Methods*, Nov. 2010. 1, 3, 6, 7

[10] Medicton group. The system I4Control. `http://www.i4tracking.cz/`. 1

[11] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, 2007. 1, 2, 5

[12] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. CVPR*, 2014. 2

[13] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Detecting driver drowsiness based on sensors: A review. *MDPI open access: sensors*, 2012. 1

[14] F. M. Sukno, S.-K. Pavani, C. Butakoff, and A. F. Frangi. Automatic assessment of eye blinking patterns through statistical shape models. In *ICVS*, 2009. 1, 2

[15] D. Torricelli, M. Goffredo, S. Conforto, and M. Schmid. An adaptive blink detector to initialize and update a view-basedremote eye gaze tracking system in a natural scenario. *Pattern Recogn. Lett.*, 30(12):1144–1150, Sept. 2009. 1

[16] X. Xiong and F. De la Torre. Supervised descent methods and its applications to face alignment. In *Proc. CVPR*, 2013. 2, 3, 4, 5

[17] Z. Yan, L. Hu, H. Chen, and F. Lu. Computer vision syndrome: A widely spreading but largely unknown epidemic among computer users. *Computers in Human Behaviour*, (24):2026–2042, 2008. 1

[18] F. Yang, X. Yu, J. Huang, P. Yang, and D. Metaxas. Robust eyelid tracking for fatigue detection. In *ICIP*, 2012. 1

[19] S. Zafeiriou, G. Tzimiropoulos, and M. Pantic. The 300 videos in the wild (300-VW) facial landmark tracking in-the-wild challenge. In *ICCV Workshop*, 2015. `http://ibug.doc.ic.ac.uk/resources/300-VW/`. 3