Regressor Based Estimation of the Eye Pupil Center

Necmeddin Said Karakoc, Samil Karahan, Yusuf Sinan Akgul

Gebze Technical University, Gebze, Kocaeli 41400, Turkey,
GTU Vision Lab: http://vision.gyte.edu.tr
{said.karakoc,samilkarahan}@gmail.com
akgul@gtu.edu.tr

Abstract. The locations of the eye pupil centers are used in a wide range of computer vision applications. Although there are successful commercial eye gaze tracking systems, their practical employment is limited due to required specialized hardware and extra restrictions on the users. On the other hand, the precision and robustness of the off the shelf camera based systems are not at desirable levels. We propose a general purpose eye pupil center estimation method without any specialized hardware. The system trains a regressor using HoG features with the distance between the ground-truth pupil center and the center of the train patches. We found HoG features to be very useful to capture the unique gradient angle information around the eye pupils. The system uses a sliding window approach to produce a score image that contains the regressor estimated distances to the pupil center. The best center positions of two pupils among the candidate centers are selected from the produced score images. We evaluate our method on the challenging BioID and Columbia CAVE data sets. The results of the experiments are overall very promising and the system exceeds the precision of the similar state of the art methods. The performance of the proposed system is especially favorable on extreme eye gaze angles and head poses. The results of all test images are publicly available.

1 Introduction

Accurate estimation of the eye pupil center locations is crucial for many applications such as eye gaze estimation and tracking [1], human-machine interfaces [2], user attention estimation [3], and controlling devices for disabled people [4]. There are several types of methods for pupil center localization. One of these types uses specialized hardware such as head-mounted devices [5] or multiple near-infrared cameras [6]. These methods estimate the center accurately but they are expensive, uncomfortable, intrusive, and generally require a calibration stage. Furthermore, the systems based on active infrared (IR) illumination are less robust in daylight applications and outdoor scenarios.

Recently, appearance based pupil center localization methods, which need only webcam type cameras, started addressing the above problems. These techniques can roughly be divided into three categories [7]: model-based methods [8] [9] [10], feature-based methods [11] [7], and hybrid methods [7] [12]. The model based methods use the holistic appearance of the eye to estimate the centers. These approaches often use classification or regression of a set of features or fitting of a learned model to estimate the location of the pupil centers [7]. Hamouz et al. [8] use Gabor filters to localize features, including eye corners and pupil centers. They generate face hypotheses in the affine space by using feature triplets. Finally, an appearance model is applied to pick out the best among the pre-selected face hypotheses. Markus et al. [10] describe a method for pupil location estimation based on an ensemble of randomized regression trees. This method employs the human body part segment classification method of MS Kinect [13] which uses difference of random pixels around the pupil center. Our proposed regression method, on the other hand, employs Support Vector Regressor (SVR) technique, which is known to be precise in terms of localization of structures [14]. Our employment of HoG [15] features also helps us take advantage of the rich image gradient angle information around the eye regions.

Feature based methods do not use any learning. They employ eye properties to detect candidate pupil centers from local image features (e.g. corners, edges, gradients). Timm et al. [11] propose an approach based on the analysis of image gradients. They define an objective function that expects the intersection of all the gradient vectors on the pupil center. Although, this method also uses image gradients like our method, our employment of gradient information involves model training which makes our system more robust for head pose and eye gaze changes.

The hybrid methods collect pupil center location candidates using feature based methods. Model based methods are employed to select the optimal values among the candidates. Valenti et al. [7] use the curvature of isophotes as image features to design a voting scheme for pupil localization. They later combine this information with the extracted SIFT [16] vectors for each candidate location and match it with examples in a database to obtain the final decision.

In this paper, we propose a new model based approach for accurate and robust pupil center localization on images supplied from a monocular camera system. We argue that the gradient angle information around the eye pupil region has a unique signature (see the extracted HoGs in Fig. 1 and Fig. 3) and employing HoG technique as the basic feature extraction tool should capture this information. Classically, HoG features are used with binary classifiers to detect objects such as pedestrians [15], which is not suitable for accurate pupil center detection. Instead of using binary detectors, given the HoG vector of an image patch, we feed this information to an SVR [17] to estimate the distance between the patch center and the pupil center. This approach eliminates overlapping positive patches problem of binary classifiers and each image patch contributes towards the position estimation of the pupil center. The estimated distances to the pupil center from image patches are used in a polynomial curve fitting approach to obtain an accurate pupil center estimation. The resulting pupil center estimation system is both very accurate and robust against extreme head positions and eye gaze angles. We employed the output of this work for an iterative pupil center refinement framework and presented the general architecture of the system in a paper [18]. Note that there are studies that employ HoG features to detect the eye regions on face image such as, [19], [20]. Our task of pupil detection uses similar HoG features but our system is based on regression methods that measures the distance to pupil center for each patch candidate which makes our system more precise.

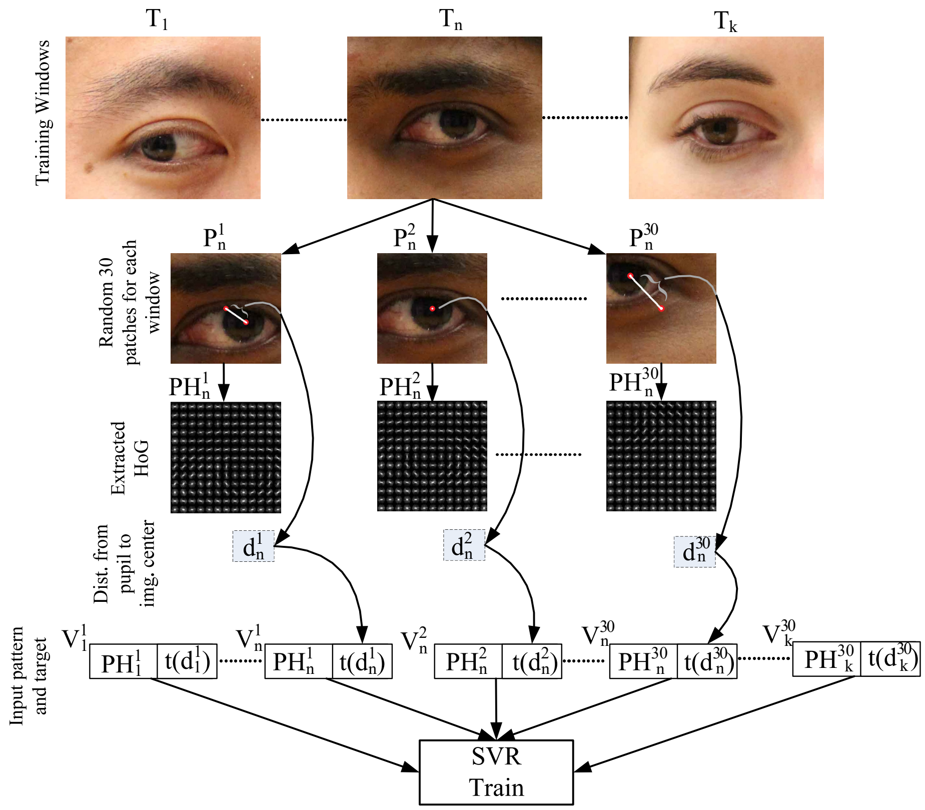

Fig. 1. Training stage of the proposed method.

The rest of the paper is organized as follows: Section 2 describes the proposed method; in Section 3 experimental results are reported and compared with the state of the art algorithms presented in the literature; Section 4 draws the conclusions of our work.

2 Method

The proposed system consists of two phases: training and testing. To train the model (Fig. 1), we use k training windows around the eye regions (T_1, T_2,…., T_k). The window sizes are chosen much larger than the eye region to make the final trained regressor robust against eye detector errors. About 30 patches (P_1^n, P_2^n, ….P_{30}^n) are randomly sampled from each T_n. The sizes of these patches are $96 \, x \, 96$.

Suppose patch P_O contains the pupil center at the patch center (such as P_2^n in Fig. 1). We like to learn a function $y(.)$ such that $y(f(P_O))$ produces maximum value where f is a function for extracting feature vectors from the image patches. The value of the function y decreases proportional to the Euclidian distance between the patch center and the pupil center. For example, $y(f(P_{30}^n))$ would produce the smallest value and $y(f(P_2^n))$ would produce the largest value for the patches of Fig. 1. We propose to learn the function y using the SVR method. To train the SVR model, we need input patterns, which are the image features produced by the function f. In our case, the function f produces HoG vectors for the given patch. SVR training also needs the targets for each input pattern. To provide this target data to the SVR model, we calculate the Euclidian distance d_r^n between the center of the patch P_r^n and the pupil center. The calculated distance values are fed to an exponential function t (Equation 1) whose values rapidly increase around the pupil centers (see the exponential function in Fig. 2.), which makes the overall localization problem more accurate [14].

$$t(d_r^n) = \begin{cases} e^{a\left(1-\frac{d_r^n}{PS}\right)} - 1 & if\ d_r^n < PS \\ 0 & otherwise, \end{cases} \quad \textbf{(1)}$$

where $a > 0$ is a constant that controls the exponential increase rate. PS is taken proportional to the patch size and it is used for producing value of zero if the distance d_r^n is larger than the patch. We observed that training a polynomial SVR model with the target values from function t (instead of d_r^n) produces much better results for the task of pupil center detection.

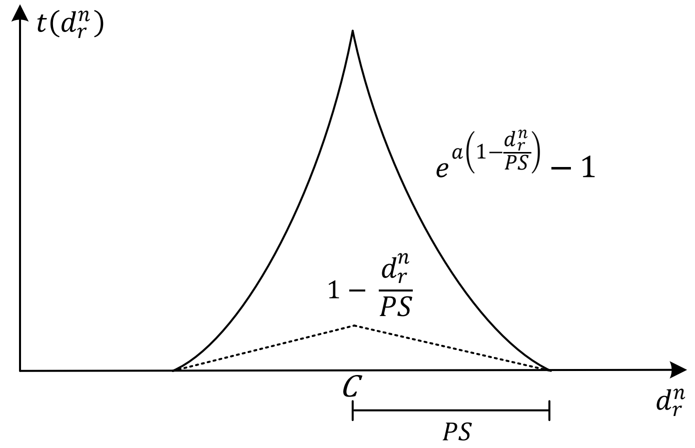

Fig. 2. The functions to generate labels against image patches.

The function f takes an image patch P_r^n and produces HoG vector PH_r^n. Finally, the vector V_r^n is formed by combining the input pattern PH_r^n with the target $t(d_r^n)$. The SVR model is trained with all V_r^n vectors, where $n = 1,2,..,k$ and $r = 1,2,..,30$. The proposed work trains only one regressor model for both eyes by flipping the left eye horizontally to make it act like the right eye. Based on our experiments, we observe that using a third degree polynomial kernel for the regressor outperforms alternative kernels.

Given a testing image I, first approximate face and eye positions are obtained (Fig. 3). There are many good face and eye detectors in the literature and our method can work with any of them such as [21]. We fix the size of the pupil center search areas (Fig. 3, E_L and E_R) proportional to the detected distance between the left and right eyes. The eye region E_L is scanned to produce sliding windows S_L^1 to S_L^m, where m is the number of sliding windows in eye region E_L. Each sliding window S_L^n is scaled to $96\,x\,96$ pixels similar to the scaling process done in the training phase. We then produce the HoG vector SH_L^n for each S_L^n. The vector SH_L^n is fed to the trained SVR to produce the estimated exponential distance to the pupil center for each S_L^n. The same process is repeated for the other eye region E_R. A score image is formed by using the regressor results which visually shows where the pupil centers are located (Fig. 3).

We employ a two stage approach for the scanning window process. First, we start from the top-left corner and slide the window by skipping 5 pixels. The regressor response is calculated for each sliding window S. The maximal response areas are then scanned again with 1 pixel window skipping. Thus, we reduce the search time considerably. For the final estimation, our method selects the best 20 positions on the score image. Second degree polynomials are fit around these maximal points on the score image. We calculate the zero value of the first derivative of the polynomials to find the pupil center. Our task of estimating the pupil centers is for both eyes, thus we run a special algorithm that finds the pupil centers of both eyes simultaneously. Out of the best 20 positions on the score image of both eyes, we chose the best combination that produces best total score within a minimum and maximum Euclidian separation. As a result, eye centers that have too close or too far positions are eliminated.

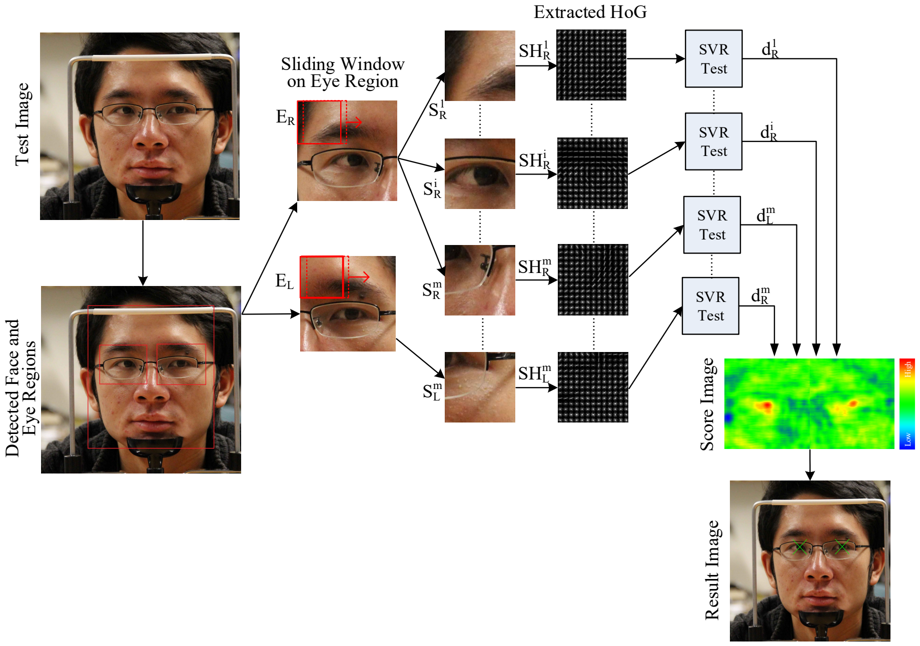

Fig. 3. Test stage of the proposed method.

3 Evaluation

For the first set of validation experiments, we have chosen the BioID database [21], which has a challenging set of images for pupil center estimation. It consists of more than 1500 gray level face images of 24 subjects taken in different locations and times, which cause variable illumination conditions. Additionally, several subjects are wearing glasses. The eyes of some subjects are closed and sometimes there are shadow effects around the eyes. In a few images, strong reflection on the glasses causes invisible pupils. The image resolution (286x384) is equal quality of a low-resolution camera. The centers of the both eyes are hand marked in the dataset. We use these eye centers as the eye detection results for this experiment. The experiments are performed using leave-3-person-out cross validation method, i.e., the system is trained with images of 21 persons, tested with 3 persons, and this process is repeated using different person combinations until all the images in the set are tested. This method guarantees no person is included both in the training and testing set. We also repeated the same experiment with 10-fold cross validation method, which might include very similar face images in both training and testing phases.

To measure the accuracy for the estimated pupil centers, the normalized error is evaluated. This measure was proposed by Jesorsky et al. [22] and is defined as

$$e \leq \frac{max(e_L, e_R)}{d} \quad (2)$$

where e_L and e_R are the Euclidean distances between the estimated and the ground-truth pupil centers for the left and right eyes, respectively. d is the Euclidean distance between the ground-truth pupil centers. Approximately, an error of $e \leq 0.25$ corresponds to distance between the pupil center and the eye corners, $e \leq 0.1$ corresponds to the diameter of the iris, and $e \leq 0.05$ corresponds to the diameter of the pupil [12]. We argue that more accurate pupil center localization methods should produce values smaller than 0.05. Table 1 compares our normalized results with the results of other known methods. As can be seen from the data, our method achieves best error rates for two categories and matches the best method for the third error category. Since the other methods did not report their cross validation details, we report our results for both leave-3-person-out and classical 10-fold cross validation. We should note that our search windows around the pupils are much larger than other methods to eliminate any problems due to eye region detection errors. If the employed eye region detectors are robust, then the search regions can be kept smaller to achieve better performance due to reduced search space. The first two rows of Fig. 4. show some of our results on BioID data set. The last column shows failure cases. The results of our system different data sets are available publicity at the project page [23].

For the second set of validation experiments, we evaluate our system on the Columbia gaze dataset, CAVE [24]. It contains high-resolution (5184x3456 pixels) images of 56 subjects each looking at 21 different positions that require a wide range of eye gaze positions. In addition to different eye gaze angles, the subject heads take one of the five head poses. We manually marked the pupil centers for 300 images, which involve nominal and extreme eye gaze angles and head poses, and report our results for 10-fold cross validation. We also run the system of [11] with the publicly available code on CAVE set. Table 2 compares our results with the results of [11]. Although [11] performs worse on this dataset due to very extreme head poses and eye gaze angles, our system takes advantage of the extra resolution to produce much better results.

Table 1. Comparison of the normalized error on the BioID set.

Method	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Our Results (leave-3- person-out)	**%92.2**	**%97.7**	%99.6
Our Results (10-fold validation)	**%94.7**	**%98.7**	**%99.7**
Markus et al. $p = 31$ **[10]**	%89.9	%97.1	**%99.7**
Tim et al. **[11]**	%82.5	%93.4	%98.0
Valenti et al **[7]** hybrid.	%86.1	%91.7	%97.9

Table 2. Comparison of the normalized error on the CAVE set.

Method	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Our Results (10-fold validation)	**%98.3**	**%99.3**	**%100**
Tim et al. **[11]**	%74.7	%78.0	%83.7

For the third set of experiments, we employed a face and eye detector to find the initial eye positions instead of the hand marked positions provided by the datasets. This is the standard practice with the other state of the art methods. The well-known face detection algorithm Viola&Jones [25] is used for this purpose. If the face and eye detector is successful, then we find the pupil centers. Otherwise, we do not estimate the pupil center. Since the face and eye detectors fail on bad images, the success rate of pupil detectors are higher for this experiment. In test stage, leave-3-person-out and classical 10-fold cross validation is followed like the first test strategy. As can be seen the Table 3, better performance is achieved than the using the full dataset. Generally, non detected face images contain the extreme head pose, profile face, or occluded scenario. These situations mostly cause to increase detection error for the first two experiments and decrease the error for the third experiment.

Table 3. Comparison of the normalized error on the detected face images in BioID set.

Method	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Our Results (leave-3- person-out)	**%97.5**	**%99.6**	**%99.9**
Markus et al. $p = 31$ **[10]**	%89.9	%97.1	%99.7
Tim et al. **[11]**	%82.5	%93.4	%98.0
Valenti et al **[7]** hybrid.	%86.1	%91.7	%97.9

In order to show the characteristics of the proposed method on nominal and extreme cases, we show the normalized error versus accuracy graph of the CAVE experiments in Fig. 5. As expected our method and [11] show similar performances as the BioID set on nominal cases. However, the performance of [11] drops significantly for extreme cases while our results stays at good levels.

The third and fourth rows of Fig. 4. show some of our results on CAVE data set. The last column shows failure cases. There are two main reasons for the failures whose score images are shown in Fig. 6. First, for some cases, our score images do not produce the correct positions of the eye pupils, which may indicate a training set problem. If the training set includes more closed eye lid examples or glass subject examples, our training could reflect the real world cases better. Second, for some cases, although the score images produce the correct positions, our simultaneous pupil estimation method fails, which suggests we may need a more sophisticated combined pupil estimation method.

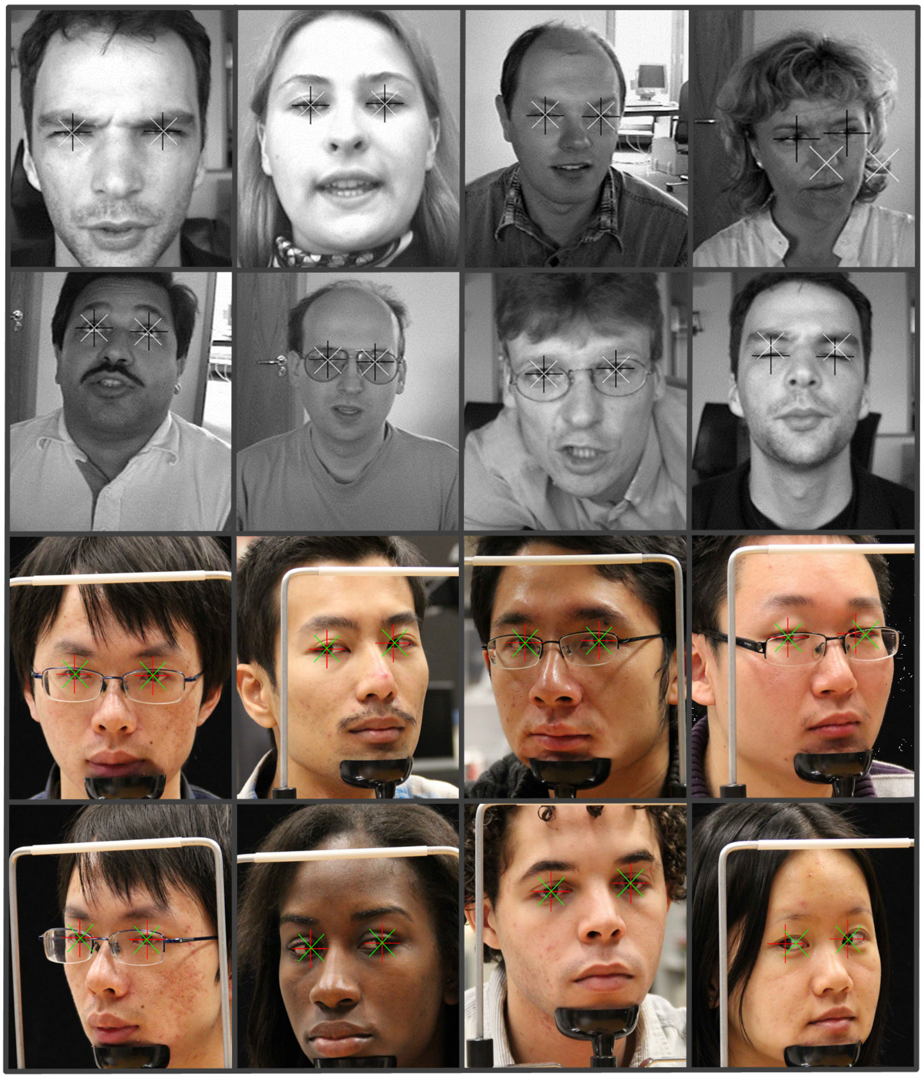

Fig. 4. First and second rows: our results on BioID, third and fourth row: our results on CAVE. The last column shows the failure cases. Plus marks show the ground truth, cross marks show the estimated pupil positions.

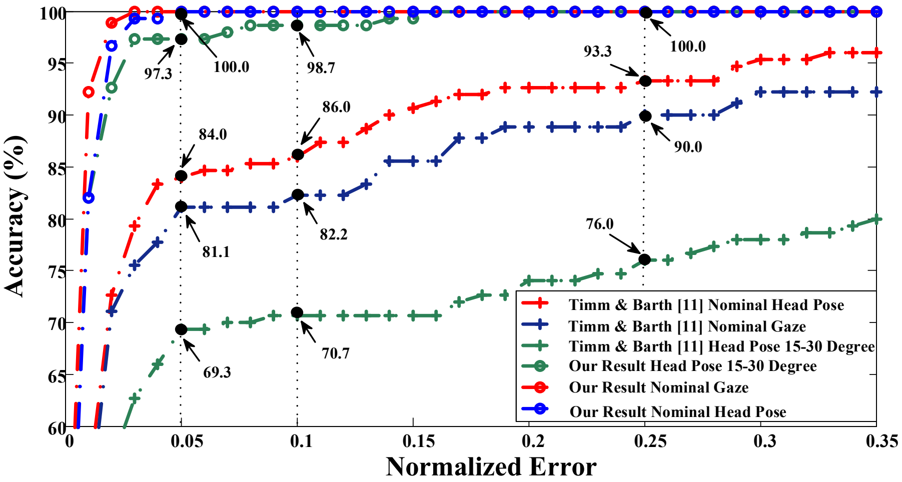

Fig. 5. Proposed method versus **[11]** on the CAVE set.

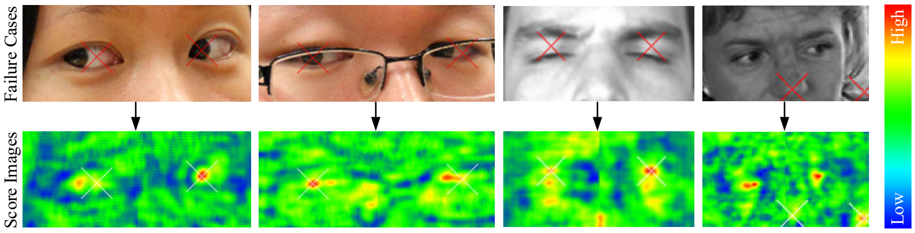

Fig. 6. First row: the failure cases, second row: score images generated from the failure cases. Cross marks show the estimated pupil positions.

4 Conclusions

We propose an appearance based approach to estimate the eye pupil centers accurately and robustly by using a Support Vector Regressor. Our method extracts HoG features from the candidate eye patches and estimates the distance of the patch center to the pupil center. The HoG features can take advantage of the gradient direction information around the eye region especially with good resolution images. As a result, our method is more suitable for pupil center localization for the eye gaze angle estimation. The experiments performed on the standard datasets show the effectiveness of the proposed method. For the future work, we plan to make our method faster by using a steepest ascent algorithm on the regressor function results that moves towards the pupil center. It is also possible to make our system faster by running parallel threads for each sliding window patch, which are independent of each other.

References

1. A. Duchowski, Eye tracking methodology: Theory and practice, Springer, 2007.
2. A. Poole and B. J. Linden, "Eye tracking in HCI and usability research," in *Encyclopedia of human computer interaction*, Pennsylvania, Idea Group, 2006, pp. 211-219.
3. K. Rayner, C. M. Rotello, A. J. Stewart, J. Keir and S. A. Duffy, "Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements," *Journal of Experimental Psychology: Applied,* vol. 7, no. 3, pp. 219-226, 2001.
4. J. Levine, "An eye-controlled computer," IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y, 1982.
5. D. Li, J. Babcock and D. J. Parkhurst, "openEyes: a low-cost head-mounted eye-tracking solution," in *Proceedings of the 2006 symposium on Eye tracking research & applications*, 2006.
6. T. Ohno and N. Mukawa, "A Free-head, Simple Calibration, Gaze Tracking System That Enables Gaze-Based Interaction," in *Proceedings of the 2004 symposium on Eye tracking research & applications*, 2004.
7. R. Valenti and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 9, pp. 1785-1798, 2012.
8. M. Hamouz, K. Josef, K. J-K., P. Pekka, K. Heikki and M. Jiri, "Feature-Based Affine-Invariant Localization of Faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 9, pp. 1490-1495, 2005.
9. P. Wang and Q. Ji, "Multi-view face and eye detection using discriminant features," *Computer Vision and Image Understanding,* vol. 105, no. 2, pp. 99-111, 2007.
10. N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg and R. Forchheimer, "Eye pupil localization with an ensemble of randomized trees," *Pattern recognition,* vol. 47, no. 2, pp. 578-587, 2014.
11. F. Timm and E. Barth, "Accurate Eye Centre Localisation by Means of Gradients," in *VISAPP*, 2011.
12. P. Campadelli, R. Lanzarotti and G. Lipori, "Precise eye localization through a general-to-specific model definition," in *BMVC*, 2006.
13. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," *Communications of the ACM,* vol. 56, no. 1, pp. 116-124, 2013.
14. A. Sironi, V. Lepetit and P. Fua, "Multiscale Centerline Detection by Learning a Scale-Space Distance Transform," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
15. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
16. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, no. 2, pp. 91-110, 2004.
17. H. Drucker, C. J. Burges, L. Kaufman, A. Smola and V. Vapnik, "Machines, Support Vector Regression," in *Advances in neural information processing systems*, 1997, pp. 155-161.
18. N. S. Karakoc, S. Karahan and Y. S. Akgul, "Iterative Estimation of The Eye Pupil Center," in *Signal Processing and Communications Applications Conference (SIU)*, Turkey, in Turkish, 2015.
19. S. Chen and C. Liu, "Precise Eye Detection Using Discriminating HOG Features," *Computer Analysis of Images and Patterns,* vol. 6854, pp. 443-450, 2011.
20. D. Monzo, A. Albiol, J. Sastre and A. A. Albiol, "Precise eye localization using HOG descriptors," *Machine Vision and Applications,* vol. 22, no. 3, pp. 471-480, 2011.

21. "BioID Image Dataset," [Online]. Available: https://www.bioid.com/About/BioID-Face-Database. [Accessed May 2015].
22. J. Oliver, K. J. Kirchberg and R. W. Frischholz, "Robust face detection using the hausdorff distance," in *In Audio-and video-based biometric person authentication*, 2001, pp. 90-95.
23. "Estimation of The Eye Gaze Direction," [Online]. Available: http://vision.gyte.edu.tr/projects.php?id=14. [Accessed July 2015].
24. B. A. Smith, Q. Yin, S. K. Feiner and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *In Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013.
25. P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision,* vol. 57, no. 2, pp. 137-154, 2004.