Learning the Deep Features for Eye Detection in Uncontrolled Conditions

Yue Wu
Dept. of ECSE, Rensselaer Polytechnic Institute
Troy, NY, USA 12180
Email: wuy9@rpi.edu

Qiang Ji
Dept. of ECSE, Rensselaer Polytechnic Institute
Troy, NY, USA 12180
jiq@rpi.edu

***Abstract*—Although eye detection has been studied for a long time in academic and industrial communities, it is still a changeling problem if facial images are with varying head poses, facial expressions, illuminations and resolution changes etc., which tend to happen in uncontrolled conditions. In this work, we propose to learn deep features that could capture the appearance variations of eyes for eye detection on those changeling facial images. Specifically, we exploit the idea of deep feature learning method, and construct eye detector based on the learned features. Experimental results on benchmark databases with different head poses, expressions, illuminations or resolutions show the effectiveness of the eye detector based on the learned features compare to state-of-the-art works.**

I. Introduction

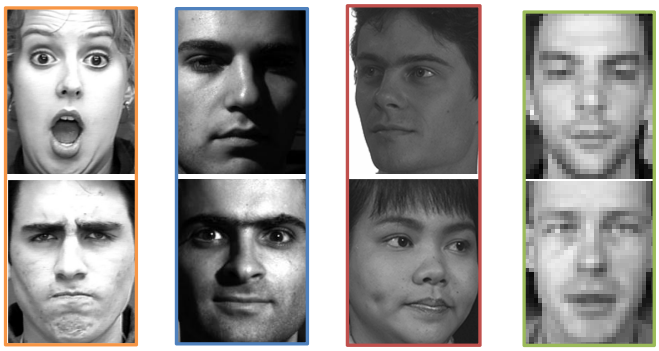
Eyes, as the most salient facial components on face, can reflect human's affective states and attention focus. The locations of eyes can provide important information for most face analysis tasks, such as face recognition and facial expression recognition. Thus, as the essential part for the success of a wide range of face-related tasks, effective and efficient eye detection algorithm has gained increasing attention in the academic and industrial communities.

A major challenge for eye detection is the large appearance and shape variations of eyes due to pose, illumination, facial expression and resolution changes. For instance, as shown in Figure 1 (a), eyes tend to open widely if facial images undergo surprised facial expression and they are half-closed with angry expression. Other examples can be viewed in Figure 1 (b)(c)(d), where eye patches vary with different poses, illuminations and resolutions.

In this paper, we propose to learn the deep features for eye detection. Specifically, the learned features should capture the large variations of eyes due to pose, expression and illumination changes, which would dramatically improve the performance of eye detection in uncontrolled conditions. Unlike traditional eye detection methods which use hand-crafted or selected features, the learned features could fit the specific eye detection task well, and thus achieve high accuracy in the real-lift scenarios with challenging uncontrolled conditions.

II. Related work

In the literature, eye detection algorithms can be classified into four categories [6][7], including the shape based approaches, the feature based shape methods, the appearance based methods, and the hybrid methods.



(a) Expression (b) Illumination (c) Pose (d) Low resolution

Fig. 1. An illustration of the sample images in various uncontrolled conditions. Images are from the (a)CK+ [1][2],(b) YaleB [3], (c) FERET [4], and (d)BioID [5] databases.

The shape based approaches exploit the distinct shape properties of the eyes, such as the circular shape of iris and elliptical shape of eyelids. For instance, in [8], Valenti and Gevers proposed method to infer the eye center locations by exploiting the circular symmetry and isophote property of the iris. Specifically, with the circular assumption, each pixel votes for its own circular center and a scale space framework is adopted to improve the accuracy. The feature-based shape methods identify a few features around the eyes such as limbus, pupil and cornea reflections and use them to support eye detection. For the appearance-based methods, they aim at modeling the variations of eye patches by statistic models. For example, in [9], Campadelli et al. proposed to train the eye detector based on the selected Haar wavelet features and the Support Vector Machines classifier. In [10], Everingham and Zisserman proposed and compared three eye detectors based on regression method, simple bayesian model and a discriminative adaboost classifier, among which bayesian model achieves the best performance. In [11], Kim et al. proposed to detect the eyes based on the multi-scale Gabor features with a coarse-to-fine strategy. The hybrid models combine at least two of the mentioned techniques.

Although these models can achieve accurate eye detection on "simple facial images", they tend to encounter problems especially on images taken in uncontrolled conditions, since eyes in these images change dramatically not only due to cross subject variations but also due to the influence of arbitrary environmental conditions (Figure 1). For example, the shape based approaches may fail if the eyes are closed or the resolu-

tion of the images is too low, since the circular shape of iris can not be viewed in these cases. Similarly, the features around the eyes cannot be robustly detected if occlusion happens, which could lead to the failure of feature-based shape methods. For most of the appearance based methods, they are based on hand-crafted features or selected features from a feature set, which may not be optimal for eye detection. Furthermore, it's difficult for them to capture the large appearance variations due to pose, expression, illumination changes. To address these issues, we proposed to learn good features that could capture the distinct patterns of eyes under varying changes for eye detection in uncontrolled conditions.

III. Feature learning for eye detection in uncontrolled conditions

In this section, we will first introduce the Deep Boltzmann Machine Model (DBM) [12], which inspires our method. Furthermore, we will discuss how to learn good deep features for eye detection. We will then discuss some important properties and benefits of the proposed method.

A. Deep Boltzmann Machine Model

Deep Boltzmann Machine (DBM) [12] is a deep symmetric graphical model with one visible layer $\mathbf{v}$ and several sequential hidden layers $\mathbf{h}^1, \mathbf{h}^2, ...$ In DBM model, the parameters θ include $\mathbf{W}^i$ $(i = 1, 2, ...)$ which captures the joint compatibility between nodes in the consecutive layers, the bias terms $\mathbf{c}$ for the visible layer, and $\mathbf{b}^i$ for the hidden layers. The joint energy function of the DBM model with two hidden layers (Figure 2) is:

$$\begin{aligned} -E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \theta) =& \mathbf{v}^T \mathbf{W}^1 \mathbf{h}^1 + (\mathbf{h}^1)^T \mathbf{W}^2 \mathbf{h}^2 \\ &+ \mathbf{c}^T \mathbf{v} + (\mathbf{b}^1)^T \mathbf{h}^1 + (\mathbf{b}^2)^T \mathbf{h}^2, \end{aligned} \tag{1}$$

With the formulated energy function, the DBM model captures the probability distribution of the visible units $\mathbf{v}$ by marginalizing over the hidden units and then normalized with the partition function $Z(\theta)$:

$$P(\mathbf{v}; \theta) = \frac{\sum_{\mathbf{h}^1, \mathbf{h}^2} exp(-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \theta))}{Z(\theta)}, \tag{2}$$

$$Z(\theta) = \sum_{\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2} exp(-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \theta)), \tag{3}$$

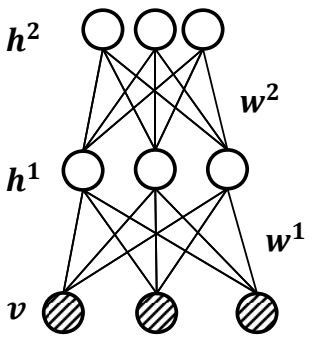


Fig. 2. Deep Boltzmann Machine model

B. Feature learning for eye detection in uncontrolled conditions

In our approach, we propose to learn the features that can capture the variations of eyes due to illumination, pose and expression changes for eye detection in uncontrolled conditions based on the Deep Boltzmann Machine (DBM) model. Specifically, the eye image patches and the background data (Figure 3) correspond to the visible input nodes $\mathbf{v}$ in the DBM model shown in Figure 2.

Given the training data $\{\mathbf{v}_i\}_{i=1}^N$, model parameters are learned by maximizing the log likelihood.

$$\theta^* = arg\max_\theta L(\theta) = arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} log(P(\mathbf{v}_i; \theta)) \tag{4}$$

The gradient of model parameters are calculated as:

$$\frac{\partial L(\theta)}{\partial \theta} = -\langle \frac{\partial E}{\partial \theta} \rangle_{P_{data}} + \langle \frac{\partial E}{\partial \theta} \rangle_{P_{model}}, \tag{5}$$

where $\langle . \rangle_{P_{data}}$ and $\langle . \rangle_{P_{model}}$ represent the expectation over the data $P_{data} = p(\mathbf{h}^1, \mathbf{h}^2 | \mathbf{v}_i; \theta)$ and model $P_{model} = p(\mathbf{h}^1, \mathbf{h}^2, \mathbf{v}; \theta)$, respectively. The data expectation is estimated based on variational approach with mean field method, while the model expectation is estimated through Markov Chain [12].

After model training, for each image patch, we could estimate its corresponding features, denoted as $\mathbf{h}^2$ in the DBM model shown in Figure 2 though the posterior probability $P(\mathbf{h}^2 | \mathbf{v}; \theta)$. Since the estimation is intractable we use Gibbs sampling method with Equation 6 and 7, where $\sigma(x) = \frac{1}{1+exp(-x)}$.

$$p(h_j^1 = 1 | \mathbf{v}, \mathbf{h}^2; \theta) = \sigma(\sum_i v_i W_{i,j}^1 + \sum_k h_k^2 W_{j,k}^2 + b_j^1), \tag{6}$$

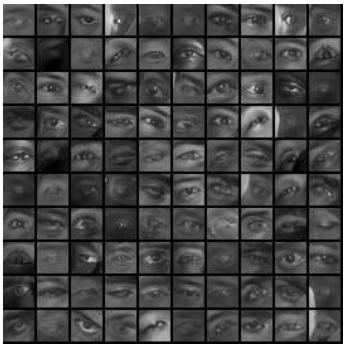
$$p(h_k^2 = 1 | \mathbf{h}^1; \theta) = \sigma(\sum_j h_j^1 W_{j,k}^2 + b_k^2), \tag{7}$$

Figure 4 shows the data in the original and learned deep feature space. As can be seen, eye patches and background patches are heavily overlapped in the original spaces, while they are separated in the feature space, which makes it easier to train a classifier to discriminatively sperate the eye image patches from the background images. Based on the learned features, we train a Neutral Network as classifier. Following a scanning window manner, eyes could be searched within the facial images.

C. Properties of the proposed methods

The proposed model has a few properties and benefits:

1) Compared to other feature learning methods, the highly nonlinearity and deep nature within the Deep Boltzmann Machine model can boost the power of the learned features to capture the image variations due to illumination, pose and expression changes, which tend to happen in uncontrolled conditions.
2) The proposed feature learning method learns task-specific features for eye detection. Unlike the hand-crafted features or selected features, they are not



(a) Eye image patches

(b) Background data

Fig. 3. Eye image patches with varying illumination, pose and expression changes, and the background data.

limited to a feature set. As a result, they have the potential to achieve the optimal features for eye detection.

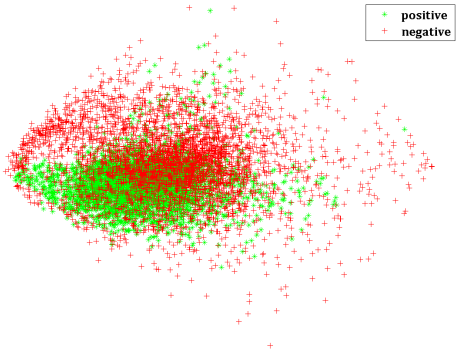
3) The proposed method does not require the clear view of iris or specific features around eyes. As a result, they could handle low resolution images or facial images with closed eyes, where most shape based and feature based shape methods tend to fail.

IV. Experimental results

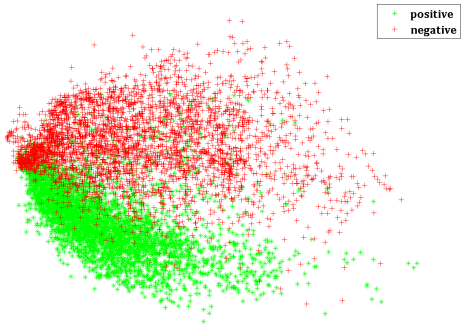
In this section, we evaluate the proposed eye detector with the learned features on several benchmark databases, including the CK+, FERET, YaleB, and BioID databases, which contain facial images with varying expressions, poses, illuminations, etc. For FERET and YaleB databases, we randomly separate the images into training and testing set with proportions of 90% and 10%, respectively. Since the CK+ and BioID contain less images, we keep the proportions as 80% and 20% for training and testing, respectively. For the following experiments, the detection error is calculated as below:

$$error = \frac{max(||D_l - L_l||_2, ||D_r - L_r||_2)}{||L_l - L_r||_2}, \quad (8)$$

where D and L represent the detected and manually labeled ground truth eyes, and the subscript denotes left and right eyes. Following the previous works [13] [9], we consider the eyes are correctly detected if $error < 0.25$. For all the experiments, we focus more on comparing our method with the appearance and image patch based methods. All the sampled image patches are normalized to the same size as input to the model.



(a) Data in original space.



(b) Data in the learned feature space.

Fig. 4. Visualize the data in the original and feature spaces (better see in color). Green points: positive eye patches. Red points: negative background data.

A. Facial expression variations

To verify our method on images with facial expression variation (Figure 1 (a)), we select the Extended Cohn-Kanade dataset (CK+) [1][2]. The CK+ database contains facial videos from 123 subjects with 7 facial expressions, including anger, disgust, fear, happy, sadness, surprised and contempt. For each sequence, we use the first, onset and apex frames. Since the images with contempt expression are very limited, we exclude them and only use the images corresponding to the other six facial expressions. In total, there are 1339 images. The eye detection results are shown in Table I. We achieve better performance than the LBP features [14] with the Viola-Jones detector [15] and the work in [13]. Please see Figure 5 for more results.

TABLE I. Eye detection results on CK+ [1][2] database.

methods	detection rates
LBP [14], Viola-Jones[15]	95.67%
Vesselness filter [13]	98.87%
Proposed method	**99.22%**

B. Pose and illumination variations

We evaluate the proposed method on the facial images with pose and illumination variations from the Extended Yale Face Database B (YaleB) [3] and the Facial Recognition Technology (FERET) [4] database.

The YaleB database contains images of 28 subjects from 64 illumination sources and 9 poses (Figure 1 (b)). To ensure

fair comparison, we follow the work in [8] and use the images where lighting sources are no more than $\pm 40^o$ (24 illuminations). In addition, we show results on three testing sets. Set (1) only contains images with varying poses without additional lighting source. Set (2) only contains images with frontal pose and varying lighting sources. Set (3) contains all images with both pose and illumination variations. As can be seen from Table II, our method achieves same or better performances than the reported results in [8] on subset (1)(2) (they don't report their results on subset (3)). It is also consistently better than the LBP features [14] with the Viola-Jones detector [15].

TABLE II. EYE DETECTION RATES ON YALEB [3] DATABASE.

Testing sets	method	detection rates
(1) Pose variation set	LBP [14], Viola-Jones [15]	93.10%
	Isocentric pattern [8]	100%
	Proposed method	**100%**
(2) Illumination variation set	LBP [14], Viola-Jones [15]	98.36%
	Isocentric pattern [8]	96.73%
	Proposed method	**100%**
(3) Illumination and pose variation set	LBP [14], Viola-Jones [15]	91.99%
	Proposed method	**97%**

The FERET database contains 4335 images with varying poses (Figure 1 (c)). There are 1417, 853 and 743 images in the frontal, quarter left and quarter right subsets, respectively. As can be seen from Table III, our method is comparable to the work in [9] and the LBP features [14] with Viola-Jones detector [15] on frontal images. It achieves better results than the Viola-Jones detector [15] on images with quarter left and right poses.

TABLE III. EYE DETECTION RATES ON THE FERET [4] DATABASE.

Testing sets	method	detection rates
(1) Frontal pose	LBP [14], Viola-Jones [15]	96%
	General-to-specific [9]	**96.4%**
	Proposed method	95.2%
(2) Quarter left	LBP [14], Viola-Jones [15]	85.48%
	Proposed method	**93.55%**
(3) Quarter right	Viola-Jones [15]	88.24%
	Proposed method	**91.18%**

C. Low resolution images

To evaluate our method on low resolution images, we perform eye detection on selected and downsampled 956 images from BioID database (Figure 1 (d)). Specifically, in the original image, face regions are with resolution 150*150. We downsample the images into 50% of their original sizes. In this case, there are only 18*18 pixels in the eye region. Table IV compares the detection rates on downsampled images and original images. As can be seen, using our method, eye detection rate on 50% downsampled images is comparable to that on the original images. The results are also comparable with those by the LBP features [14] with Viola-Jones detector [15].

TABLE IV. EYE DETECTION RATES ON THE BIOID DATABASE [5].

Image resolutions	face size	eye size	methods	detection rates
Original image	150*150	36*36	LBP [14], Viola-Jones [15]	98.64%
			Proposed method	98.12%
50% downsampling	75*75	18*18	LBP [14], Viola-Jones [15]	98.01%
			Proposed method	98.12%

V. CONCLUSION

In this paper, we have proposed an eye detector based on learned deep features for detection in uncontrolled conditions. Specifically, the learned deep features can capture the large appearance variations of eye patches on images with varying head poses, facial expressions, illuminations and resolutions. We show the effectiveness of the eye detector on several benchmark databases, including FERET, YaleB, CK+, BioID. Comparing to several state-of-the-art works, our detector can deal with more changeling images and achieve similar or better accuracies. In the future, we will improve the model and extend the experiments on more changeling images. In addition, we will compare our method with more latest appearance and image patch based methods, and other methods.

ACKNOWLEDGMENT

This work is supported in part by a grant from US Army Research office (W911NF-12-C-0017).

REFERENCES

[1] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, March 2000, pp. 46 – 53.

[2] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression," in *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis*, 2010, pp. 94 – 101.

[3] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[4] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295 – 306, 1998.

[5] "Bioid database. http://www.bioid.com/index.php?q=downloads/software/bioid-face-database.html."

[6] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[7] F. Song, X. Tan, S. Chen, and Z. Zhou, "A literature survey on robust and efficient eye localization in real-life scenarios," *Pattern Recognition*, vol. 46, no. 12, pp. 3157 – 3173, 2013.

[8] 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.

[9] P. Campadelli, R. Lanzarotti, and G. Lipori, "Precise eye localization through a general-to-specific model definition." in *BMVC*, 2006, pp. 187–196.

[10] M. Everingham and A. Zisserman, "Regression and classification approaches to eye localization in face images," in *International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 441–446.

[11] S. Kim, S. Chung, S. Jung, D. Oh, J. Kim, and S. Cho, "Multiscale gabor feature based eye localization," *Proceedings of world academy of science*, vol. 21, pp. 483–487, 2007.

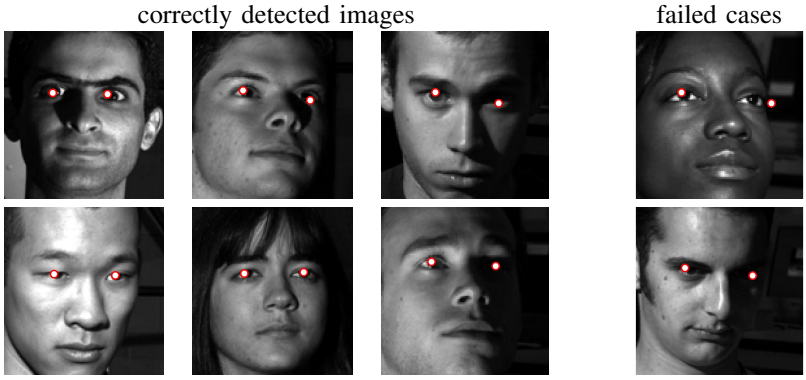
[12] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 5, 2009, pp. 448–455.

[13] A. Poursaberi, S. Yanushkevich, and M. Gavrilova, "Modified multi-scale vesselness filter for facial feature detection," in *Fourth International Conference on Emerging Security Technologies (EST)*, 2013, pp. 21–24.

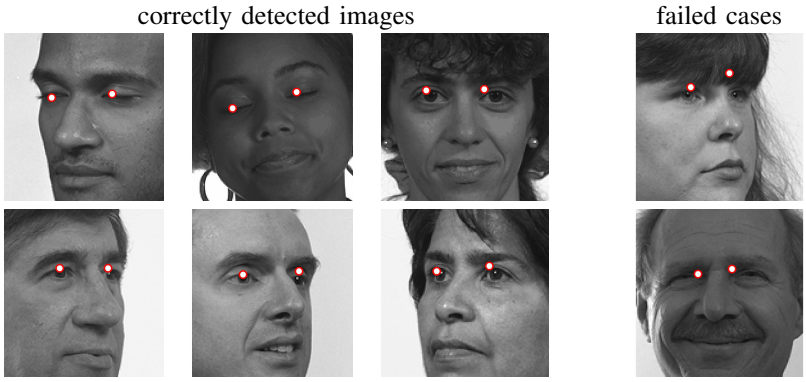
[14] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, July 2002.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference*, 2001, pp. 511–518.

(a) CK+ database [1][2]



(b) YaleB database [3]



(c) FERET database [4]

(d) BioID database [5]

Fig. 5. Eye detection results on different databases.