Pattern Recognition 98 (2020) 107076

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Eye landmarks detection via weakly supervised learning

Bin Huang*, Renwen Chen, Qinbang Zhou, Wang Xu

State Key Laboratory of Mechanics and Control of Mechanical Structures, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 7 January 2019
Revised 15 September 2019
Accepted 7 October 2019
Available online 9 October 2019

Keywords:
Eye landmarks detection
Special format data
Weakly supervised learning
Object detection
Recurrent learning module

ABSTRACT

Extensive eye researches provide good results when images are captured under constrained environment. However, the accuracy of eye landmarks detection depends on explicit bounding-box of eye regions and drops severely in non-ideal conditions. This paper has proposed a novel weakly supervised eye landmarks detection algorithm with object detection and recurrent learning modules. The former is combined with faster R-CNN and is competent to detect bounding-box of facial components and initial positions of the eye. The recurrent module is employed for eye landmarks refinement using the initial eye shape. The proposed algorithm can augment training data effectively and our specific format data consist of supervised and weakly supervised samples. Supervised samples have ground truth of bounding-boxes, corresponding classification labels and eye landmarks coordinates while weakly supervised data does not have eye landmarks information. Despite trained on facial images, the proposed method can detect eyes in severely occluded or local view of facial images without prerequisites of face alignment. Further experiments are performed on our supervised testing set and some public datasets. Their results demonstrate the robustness and effectiveness of the proposed method.

1. Introduction

Eye localization or iris extraction is a significant technique for face recognition, iris recognition, eye tracking, expression understanding, behavior recognition etc. Particularly in terms of fatigue driving detection, eye positions and motion characteristics are the foundation of extracting fatigue-related features, such as gaze detection, blink frequency, percentage of eyelid closure time and so on [1]. Eye detection under non-ideal conditions is still an existing problem, because the accuracy of eye detection method is inevitably influenced by illumination, poses, occlusion, individual differences, image resolution and so on. Recently, deep CNNs (Convolutional Neural Networks) improve the state-of-the-art performance in many computer vision tasks, but CNNs are limited in eye researches due to the lack of large-scale datasets.

Eye landmarks detection has achieved great success in recent years. Generally speaking, traditional eye localization methods can be categorized into three types, including shape-based [2–5], appearance-based [6–9], and hybrid models [10,11]. Though these countermeasures could basically deal with one or several noise factors, their detection performance would decrease in non-lab environment. With some large eye datasets established, deep learning methods have many excellent works of eye detection [12–15]. Meng et al. [12] use five points to represent the eye shape and they build a neural network like Lenet [16]. Experimental results show that their model can detect the movement of the pupil. Krafka et al. [13] build GazeCapture, a mobile-based eye tracking dataset captured from 1450 people. They train deep convolutional neural networks based on their dataset and achieve a big improvement over previous algorithms. However, their method needs extra input such as left and right eyes. Zhang et al. [14] build an in-the-wild eye dataset by natural everyday laptop. Their CNN-based approach also achieves state-of-the-art performance of eye detection, which is trained on eye images represented by six facial landmarks. Their method might fail due to inaccurate positions of eye contour. In another work proposed by [15], the model of two-level cascaded CNNs is proposed to solve eye landmarks detection in eye state conversion between open and close. They also build an eye dataset collected from the Internet. However, their methods also request to locate eye patches in facial images by using face alignment algorithm, which would bring in redundant information and affect accuracy of eye landmarks detection. If facial key points fail to be located, then eye landmarks detection is impossible. Besides, these deep learning methods usually build a large dataset of eyes, which are generally difficult or time-consuming to collect numerous labeled samples or to calibrate them manually.

In this paper, we propose a weakly supervised eye landmarks detection algorithm to locate eye points directly in a whole or local view of facial image. In our implementation, we first build an eye dataset including supervised and weakly supervised samples, which come from initial public datasets. Supervised samples

* Corresponding author.
E-mail address: binhuang@nuaa.edu.cn (B. Huang).

https://doi.org/10.1016/j.patcog.2019.107076
0031-3203/

Table 1
Comparison of our dataset with public datasets. "People" indicates the number of participants whose images are collected. "Ill." is the abbreviation for illumination. We use "greatly" represent there is continuous variation in pose or illumination. "Source" means the equipment for capturing images, which is related to image quality.

	People	Ill.	Pose	No. of images	Source
PoseGaze [20]	20	1	19	2220	Digital camera
GazeLocking [21]	56	1	5	5880	Digital camera
Eyediap [22]	16	2	greatly	94 videos	HD camera (lab)
TabletGaze [23]	51	greatly.	greatly	816 videos	Tablet front camera
Ours	5270	greatly	greatly	57,674	Public dataset

have three types of ground truth. One is classification label of bounding-box for each facial component; the second is coordinates of bounding-box for each facial component and the last is coordinates of eye-related landmarks. Weakly supervised samples do not have information about eye landmarks. We also make a few optimizations and rules to our fused samples in order to improve the performance of our algorithm. Using our specific sample format, the proposed method creatively incorporates with object detection method under the paradigm of weakly supervised learning. We construct our model by utilizing faster R-CNN [17] and recurrent learning modules. Faster R-CNN module can simultaneously detect bounding-box of facial components and coordinates of eye landmarks, and then recurrent learning module fine tunes positions of eye points with input of initial eye shape from faster R-CNN. Two experiments show that the two sub-modules can enhance the performance of eye landmarks detection. Besides, we conduct several comparing experiments for both facial landmarks localization and eye landmarks detection. Finally, our dataset will be made public for academic and research purposes.

Compared to eye-patch-to-landmarks methods, the proposed methods have several advantages, for example:

1) The proposed method is not dependent on the facial landmarks detection method chosen and the scale of cropped eye patch.
2) The proposed method can work even on facial images that are severely occluded or local view of facial images with eyes when facial landmarks detection method fails.

The rest of the paper is organized as follows. The following section reviews related work on eye localization algorithms. Section 3 presents the process of building our dataset and describe the details of weakly supervised eye landmarks detection algorithm including faster R-CNN and recurrent learning modules. Experiments on our own test set, UBIRIS.v2 [18] and MICHE [19] databases are the topic for Section 4. Finally, some conclusions and future work are discussed in Section 5.

2. Related work

2.1. Eye detection methods

Traditional eye detection methods mainly consist of three categories, which are referred to as appearance-based methods, shape-based methods and synthesis approaches. Appearance-based methods generally rely on template matching, which constructs an eye image model and then detect eyes using similarity calculation. Hallinan [6] builds an eye model by detecting intensity valleys and peaks and minimizes the energy function to fit the test images. Vijayalaxmi and Rao [7] utilize Gabor Filter to represent eye/non-eye patterns and train a shallow neural network as the eye classifier. Their model could deal with illumination changes and rotations of small angles. However, this method ignores the spatial information of the image and thus is hard to handle translation variance. Huang and Wechsler [8] use optimal wavelet packets for capturing the most significant characteristics of eye regions and radial basis functions for classifying facial areas as eye regions or not. Wang et al. [9] propose the nonparametric discriminant analysis method for eye detection. The appearance-based methods usually need to collect a large amount of training data of eyes, under different illumination and head poses.

Shape-based methods use a prior model of the eyelid, iris and pupil shape to match the test samples. Young et al. [2] use Hough transform to effectively extract iris and pupil that are regarded as elliptical contours. Lam and Yan [3] extend the deformable template to parameterize eyelids and iris represented by two intersecting parabolas and a circle. Furthermore, they add four corners locating at intersections of eyelids and iris. However, their method requires highly initial positions of the eye template. Kawato and Tetsutani [4] utilize a circle frequency filter to detect candidate points "Between-the-Eyes" and a rotation angle in the image plane. Eyes are detected as small darkest parts on each side of detected point.

Synthesis approaches combine shape and appearance methods to develop their respective advantages. Xie et al. [10] propose a part-based model, which uses a shape model for the location of several subcomponents and models the appearance implicitly. Ishikawa et al. [11] employ an AAM (Active Appearance Model) to fuse shape and appearance models. The shape representation is built on ASM (Active Shape Model) while the shape-independent texture is modeled by Principal Component Analysis. Their model can detect and track the eyes simultaneously.

Compared with these eye landmarks detection methods, our work can be regarded as a data driven appearance-based model instead of relying on a good initialization of eye template. We creatively incorporate eye landmarks detection with object detection method under the paradigm of weakly supervised learning. Hence, our model benefit from the powerful features learned by the CNNs to better locate the eye positions from facial backgrounds, and recurrent learning module help adjusting positions of eye landmarks.

2.2. Eye datasets and deep models

There are a few public datasets for eye detection research [20–23]. Table 1 reports some differences between these datasets and our dataset. Datasets in [22–22] have major constraints that they are recorded under controlled environment, i.e. fewer variations in head poses or lighting conditions. Although dataset in [23] contains continuously variations in head poses and illumination conditions, collecting in an indoor laboratory setting leads to limited variability of eye appearance. We consider this by using facial images in the wild, where faces have large variations in poses and illumination. While the number of people in [20,22] is rather limited, which might cause poor performance of eye landmarks detection on new facial images. We alleviate this limitation by greatly collecting images from many participants. Each one samples several images with different poses and expressions. Besides, samples in our dataset are generated and recalibrated from some public datasets. This way can avoid time-consuming and laborious to

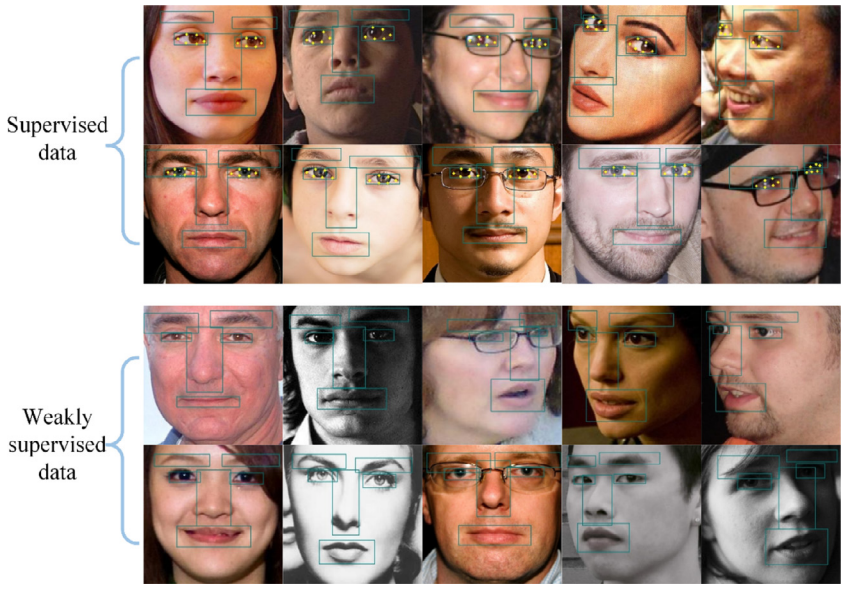

Fig. 1. Some samples from our dataset. The supervised data have bounding-box related (the ground truth labels and coordinates of each facial component) and eye landmarks coordinates inside the bounding-box of eye, while the weakly supervised data just contain the former.

collect a large number of images. We would describe the process of generating special format data in Section 3.1.

With the great improvement of deep learning made in the domain of computer vision, many deep neural networks have been applied to eye-related tasks. The earliest utilization of neural network to solve eye detection is proposed by Reinders et al. [24]. They present a multilayer perception method to locate eyes within frontal view face image without the need of any manually designed features. Kyle et al. [13] take facial images patch and facial mask as inputs to learn joint CNNs, and consequently combine this information to infer the location of gaze. Huang et al. [15] propose a multi-task learning algorithm for eye state estimation and eye landmarks detection. They use coarse-to-fine methods, i.e. two-level cascaded CNNs to fine-tune the eye shape. In our implementation, our model can also be regarded as a multi-task learning architecture, i.e. we use faster R-CNN to achieve facial component patches detection (eyebrow, eye, nose and mouth) and eye landmarks detection, simultaneously.

Recently, there are various thoughts to improve landmark localization, e.g. 3D face models [25,26], GANs [27,28], unsupervised learning [29,30], etc. Xiao et al. [25] propose recurrent 3D-2D dual learning model which alternately refines 3D face model and 2D landmarks. Bulat and Tzimiropoulos [28] employ face alignment network which is combined and jointly trained with a WGAN-based super-resolution network. Their model produces accurate landmarks localization in low resolution images. Dong et al. [30] propose the supervision-by-registration framework to improve the performance of landmarks localization. Their method brings more supervised information to enhance facial landmarks detector by using abundant unlabeled videos in the paradigm of unsupervised learning. Our method is the first attempt to explore weakly supervised learning technique combined with object detection for eye landmarks detection. In other words, we utilize special format data to achieve weakly supervised eye landmarks detection by using faster-RCNN. Importantly, our model focus on eye patches and can work with local view of facial image without face alignment methods. This phenomenon would be seen in Section 4.3.

Recurrent learning module has improved the state-of-the-art performances in speech recognition [31] and machine translation [32] applications. And many researches of key points detection [33,34] also utilize recurrent module to improve detection accuracy. Trigeorgis et al. [33] propose a combined convolutional recurrent neural network architecture for face alignment and they achieve their end-to-end training for fine tuning the face shape. As we know, we are the first to attempt the recurrent module to fine tune the eye shape based on detection results of faster R-CNN. The LSTM (Long Short-Term Memory) proposed by [35] with shared parameters would fuse spatial middle stage information for gradually adjusting positions of eye landmarks.

3. The proposed semi-supervised learning algorithm

In this section, we describe the architecture of weakly supervised eye landmarks detection algorithm. First, we introduce the pipeline for our dataset building. Then we shed light on the overview of the proposed architecture, including faster R-CNN and recurrent learning modules.

3.1. Dataset building and modification

For simplicity, we use some public annotated datasets LFPW [36], Helen [37], AFW [38], IBUG [39] and AFLW [40] to build our dataset. We would give a detailed description on these datasets. LFPW is used to evaluate facial key points detection methods. It initially consists of 1132 images for training set and 300 images for testing set. But we just download 811 training images and 224 testing image from the ibug website. HELEN contains 2000 training images and 330 testing images with a highly detailed annotation. AFW is generally used to train models, which has 337 images. IBUG only contains 135 images with large poses and expressions. AFLW has about 25k annotated facial images, exhibiting a large variety in appearance and natural conditions.

Each image is cropped out of an amplified face, which contains 68 annotated points. As the same way, the bounding-box of four facial components is defined as the max region of annotated points with appropriate amplification. Then we adopt horizontally flipping to augment our dataset. Finally, we get 57,674 training facial images, including 7674 supervised data and 50,000 weakly supervised data. We present some samples from our dataset in Fig. 1. We preserve all bounding-boxes of facial components, e.g. eyebrow, eye, nose and mouth. We think shape constraints among individual

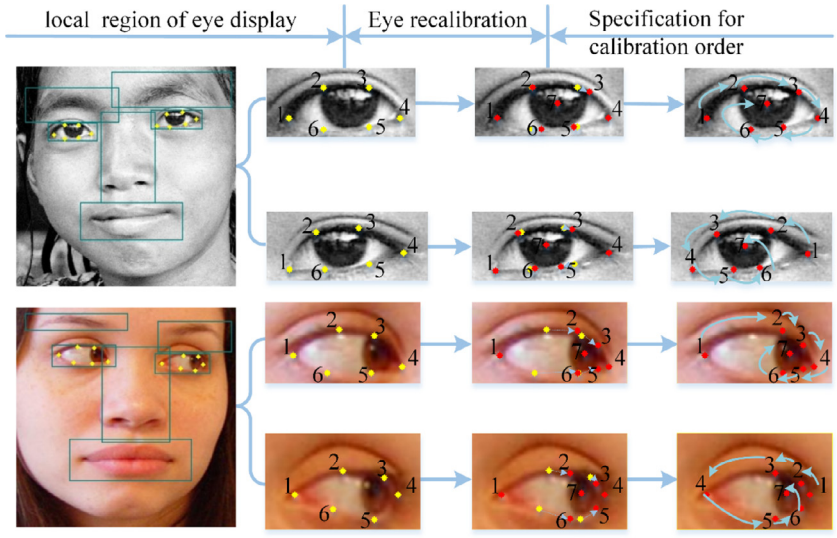

Fig. 2. Process for eye recalibration. The yellow mark indicates eye landmarks of public datasets, while the red mark indicates results of eye recalibration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

facial parts are implicitly encoded during model learning, which facilitates eye patches detection. As for all annotated landmarks, the points around eyebrow, nose, mouth and facial contour are redundant for eye landmarks detection. To reduce computational burden and force our model paying more attention to eye regions, we just keep eye-related points and ignore the rest. Nevertheless, since our samples are originally collected to study face alignment tasks, we find that they might not be suitable for precise localization of eye shape. Therefore, we make a couple of optimizations and rules to our dataset.

As showed in Fig. 1, faces have glasses occluded in column 3 and vary on different illuminations (column 2) and poses (column 4). The last column of Fig. 1 presents more complicated samples with composite noises in dataset. Further observed in Fig. 1, regardless of the position of iris in the eye, the right and left corners of the eye are represented by one point, respectively, and the remaining four points of eyelids are located at three equal points of upper and lower eyelids. The second column of Fig. 2 is magnified local patch of eye image, which could show more details about eye landmarks. It is clear that the punctuation method for the shape of eye is not sensitive to the position of iris. Therefore, in analogy with requirements of the deformable eye model, the remaining two points of eyelids are the intersection of eyelids and iris. Besides, the position of iris is not accurate to estimate the average of points around the iris due to saccadic eye movements. So we add an extra point to indicate the position of the iris, which is located at the centre of the iris.

Finally, we would like to normalize the order of calibrations for left and right eye. As seen in the third column of Fig. 2, both right and left eye regions are calibrated clockwise, which would bring in the obstacle of in network learning. It is obvious that point 1 is located at the outer Canthus of the right eye image, while point 1 is at the inner Canthus of the left eye image.

In Fig. 3, we plot a t-SNE (t-Distributed Stochastic Neighbor Embedding) [41] visualization of SIFT (Scale Invariant Feature Transform) [42] features for 2000 selected outer and inner Canthus patches of eye. These parts are centered on the left or right corner of eyes, whose size is 32×32. T-SNE is capable of providing each sample point a location in a two or three-dimensional map with the dimension reduction of features. As shown in Fig. 3, the distribution of outer and inner Canthus is approximately linearly separable. Therefore, for the left eye, we take the left corner as the

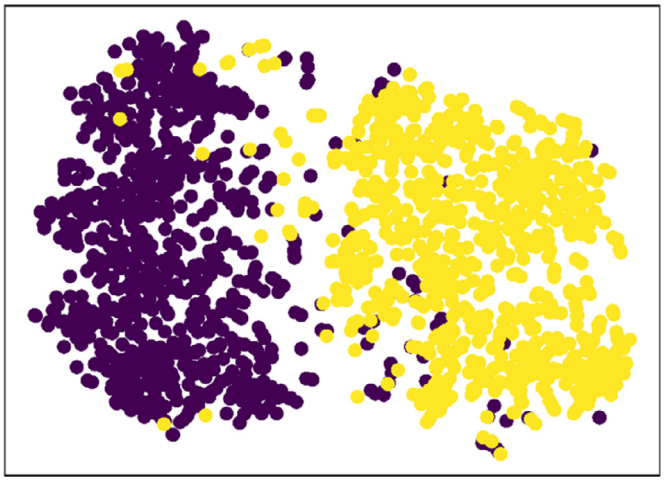

Fig. 3. A t-SNE depiction of SIFT features when asked to align 2000 randomly selected outer and inner Canthus patches of eye. The yellow circle indicates outer Canthus patches of eye, and the purple circle indicates inner Canthus patches of eye. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

starting point and calibrate the rest points clockwise; for the right eye, we start with the right corner and calibrate the rest points anti-clockwise. The fourth column of Fig. 2 gives final results of eye calibration.

To obtain ideal data, four researchers recalibrate the eye shape independently and then we take the average of their records as ground truth data.

After doing eye recalibrations on the eye regions of facial image, we obtain eye landmarks estimation dataset. Let $S^i = (x^i_{g1}, y^i_{g1}, \ldots, x^i_{g7}, y^i_{g7})$ denotes the ground truth of eye landmarks position for the i th image. Given a facial image I, we utilize RPN (Region Proposal Networks) generating detection proposals. When the region proposals have IoU (Intersection over Union) overlap with any ground truth bounding-box of facial component in the interval [0.5, 1.0], they would be regarded as positives, i.e. their label $c \in \{0,1,2,3,4\}$ is ≥ 1. Similarly, we treat region proposals with IoU in the interval [0.1, 0.5) as negatives and they are labeled with $c = 0$. For bounding-box of region proposals, the parameterization

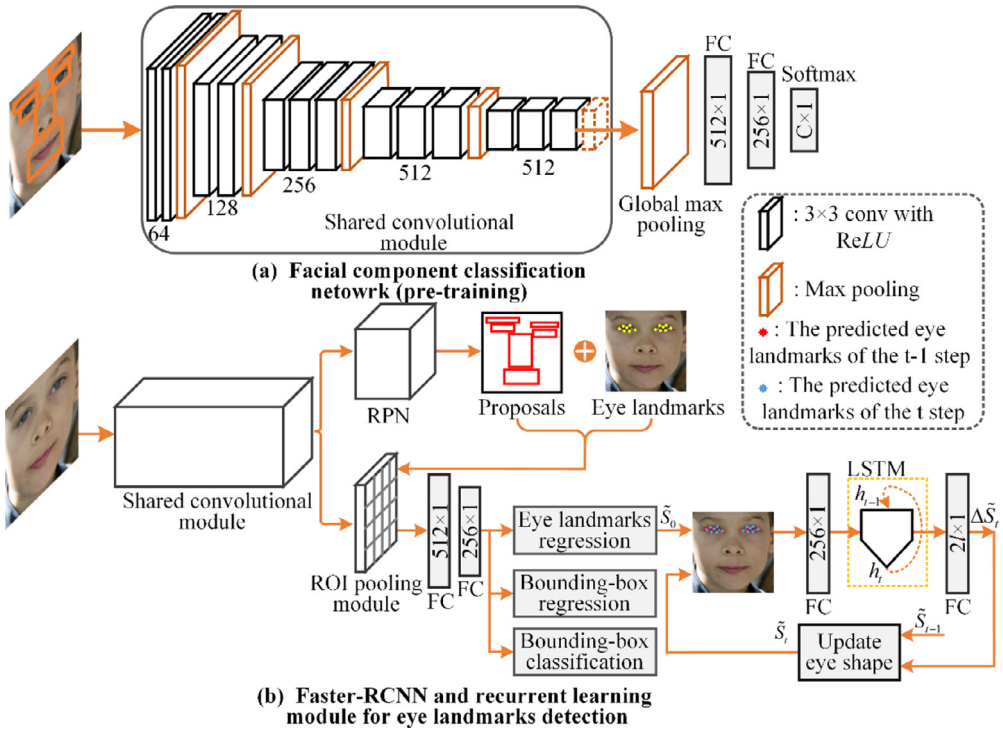

Fig. 4. Architecture of weakly supervised eye landmarks detection algorithm. The input of (a) is facial components with orange bounding-box. The input of (b) is the whole facial image. 2×2 max-pooling windows are applied in all pooling layers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

method given in [43] is applied to convert their coordinates related to ground truth bounding-box, as shown in Eq. (1).

$$\begin{cases} t_x = \frac{g_x - r_x}{g_w}, t_w = \log\left(\frac{g_w}{r_w}\right) \\ t_y = \frac{g_y - r_y}{g_h}, t_h = \log\left(\frac{g_h}{r_h}\right) \end{cases}. \tag{1}$$

where $r = (r_x, r_y, r_w, r_h)$ denotes coordinates of bounding-box for region proposals, $g = (g_x, g_y, g_w, g_h)$ denotes coordinates of ground truth bounding-box for any facial component, and the output $t = (t_x, t_y, t_w, t_h)$ denotes bounding-box offsets of region proposals related to g.

Besides, we convert eye landmarks coordinates $l = (x_1, y_1, \ldots, x_7, y_7)$that region proposals belong to eye labels, which is formulated as follows:

$$S_{xi} = \frac{x_i - g_x}{g_w}, S_{yi} = \frac{y_i - g_y}{g_h}, i = 1, \ldots 7 \tag{2}$$

where $S = (S_{x1}, S_{y1}, \ldots, S_{x7}, S_{y7})$ is relative coordinates of ground truth landmarks with regard to bounding-box of eye region. Overall, we have done all preprocessing for our dataset and we would train our model in Section 4.

3.2. Analysis of overall architecture

As shown in Fig. 4(b), the proposed weakly supervised eye landmarks detection method is roughly separated into two sub-modules: one is faster R-CNN for detecting bounding-box of facial components and coordinates of eye landmarks from facial images, and the other is recurrent learning module for fine adjusting positions of eye landmarks.

We configure faster R-CNN using VGG-16 (Visual Geometry Group) [44] as the backbone, which has 13 shareable convolutional layers. We take facial images, the corresponded proposal regions and eye-related landmarks as inputs. Faster R-CNN can be viewed as multi-task learning with three outputs, i.e. the predicted classification label, the predicted coordinates of bounding-box and the predicted coordinates of eye-related landmarks. The classification label represents the category of region proposal, which belongs to background, eyebrow, eye, nose or mouth. The coordinates of bounding-box are the offsets between ground truth bounding-box and the input region proposal. The coordinates of eye-related landmarks indicate the offset coordinates of eye landmarks related to the region proposal of the eye. The relevance between eye landmarks detection and facial components localization forces our model to extract better representations for eye shape.

For a training batch with M training region proposals as denoted by $\{I^i, c^i, t^i, S^i\}_{i=1}^M$, we can optimize faster R-CNN's parameter θ_c as follows:

$$\theta_c = \arg\min_{\theta_c} L\left(I^i, c^i, t^i, S^i, f_{RCNN}(\theta_c)\right) \tag{3}$$

where i denotes the i th region proposal and c^i, t^i and S^i are ground truth of classification label, bounding-box and eye-related landmarks, respectively. $f_{RCNN}(\theta_c)$ outputs corresponding predicted label $\widetilde{c^i}$, bounding-box $t^i$and relative coordinates of eye landmarks $\widetilde{S^i}$of the i th region proposal. To achieve training of our model, a multi-part loss function L is defined in Eq. (4):

$$\begin{aligned} L(\theta_c) = \lambda_c \sum_{i=1}^{M} H_{cls}(\widetilde{c^i}, c^i; \theta_c) + \lambda_l \sum_{i=1}^{M} 1_{[c\geq 1]}(c^i) L_{box}(\widetilde{t^i}, t^i; \theta_c) \\ + \lambda_S \sum_{i=1}^{M} 1_{[c=2]}(c^i) L_2^2(\widetilde{S^i}, S^i; \theta_c) \end{aligned} \tag{4}$$

where $H_{cls}$is the categorical cross-entropy function for classification task of the i th region proposal. $L_{box}$is a robust $L1$ loss for bounding-box regression task of the i th region proposal, which is defined in Eq. (5). The indicator function $1_{[c\geq 1]}(c^i)$ equals to 1 when $c \geq 1$, which acts as a switch that turns off the bounding-box loss when the i th region proposal is background. L_2^2 is the

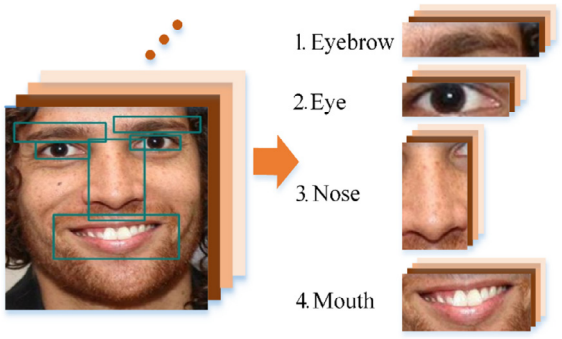

Fig. 5. Creation of facial component dataset. The facial part of distinct class is resized to same size. Each facial patch contains its facial index for classification.

Euclidean distance among seven landmarks of eye shape.

$$L_{box}(\tilde{t^i}, t^i) = \sum_{j\in(x,w,w,h)} smooth_{L1}(\tilde{t_j^i} - t_j^i),$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, |x| < 0.5 \\ |x| - 0.5, otherwise \end{cases} \quad (5)$$

The hyper-parameters λ_c, λ_l, λ_S should balance the effect of each loss part. There are two reasons to illustrate hyper-parameters setting for training. First, positives and negatives are randomly selected with a rate of 1:3 in a mini-batch and positives are evenly sampled with each facial component. Second, we should guarantee that our model focus on eye landmarks detection task. Therefore, we set $\lambda_c = 1/3$, $\lambda_l = 1$, $\lambda_S = 3$ in our training schedule.

For training our model, it is not reasonable to directly use pre-trained ImageNet VGG-16 model. Our model with eye patches localization and eye landmarks detection tasks requires fine spatial localization. The task gap between our model aiming to facial components and ImageNet classification task may limit the benefits of pre-training. Therefore, we intend to train feature extraction network as shown in Fig. 4(a). The last pooling layer in VGG-16 is replaced with global maxing pooling to eliminate the requirement of fixed-size input to CNNs. A facial patch dataset is shown in Fig. 5. Each facial component has different aspect-ratios (e.g. eye and nose).

3.3. Recurrent learning module

After faster R-CNN module finishes facial components localization and eye landmark detection, we obtain initial positions of eye landmarks for recurrent learning module. We adopt recurrent learning module to improve the performance of eye points detection based on two considerations. First, eye landmarks detection is a highly semantic task, which is mainly affected by facial appearance, illumination and pose, akin to other shape regression tasks. Second, the input for faster R-CNN module is the whole facial image, which is large but has irrelevant regions to precisely locate positions of eye points. Therefore, we take the initial eye points of outputs from faster R-CNN as input and design recurrent module to fine-tune the positions in eye regions.

We extract the initial relative coordinates of eye landmarks from the branch of eye landmarks regression, which is represented by $\tilde{S_0}$ in recurrent learning module. As shown in Fig. 4(b), the recurrent module consists of LSTM layers and two full-connected layers, which is trained end-to-end by using a single network. The corresponding full-connected layers and LSTM layers shared weights of different stages. By utilizing the LSTM layers, the middle level representation in a deep network brings useful information and can be merged well for shape estimation of the next stage. This approach also facilitates to jointly optimize different stages by extracting the middle level information.

More specifically, we optimize the recurrent network's parameters θ_r as follows:

$$\theta_r = \arg\min_{\theta_r} L\left(I_{eye}^i, S^i, S_0^i, \theta_r, T\right)$$
$$= \arg\min_{\theta_r} \sum_{t=1}^{T}\sum_{i=1}^{N_{eye}} L_2\left(S^i - S_{t-1}^i, f_R\left(I_{eye}^i, S_{t-1}^i, h_{t-1}^i, \theta_r\right)\right) \quad (6)$$

where I_{eye}^i indicates the eye region and N^{eye} indicates the total number of eye regions. T is the number of time steps. f_R outputs eye shape increment and LSTM use feedback connections to observe the middle level features $\{h_t; t = 1, \cdots, T\}$, which is formulated in Eq. (7). Note that h_t is related to $(I_{eye}, S_{t-1}, h_{t-1}, \theta_r)$. h_t would take advantage of middle level feature and eye positions of all previous stages.

$$h_0 = 0, h_t = p(I_{eye}, S_{t-1}, h_{t-1}, \theta_r) \qquad t = 1, ..., T \quad (7)$$

4. Experiments

This section is intended to demonstrate the efficiency and accuracy of the proposed weakly supervised eye landmarks detection algorithm. Then we compare the proposed method with state-of-the-art methods about performance of facial landmarks localization on LFPW and 300 W datasets. Finally, our method is compared with other methods about the performance of eye detection on UBIRIS.v2 and MICHE databases. The configuration of our model is as follows.

Datasets. We collect supervised samples using several publicly in-the-wild datasets including LFPW, Helen, AFW and IBUG. In our experiment, the 68-points annotation are used for our dataset building, which is able to be downloaded from the ibug website. As the preprocessing step, face bounding-boxes are provided by the datasets. Supervised samples for training set come from LFPW, HELEN training set and the entire AFW. The left of supervised samples belongs to testing set. Besides, we use AFLW to set up our weakly supervised dataset for training schedule, because the 68-points annotation of AFLW are obtained by computer rather than manual annotation and thus coordinates of some landmarks may be not accurate as other datasets.

Hyper-parameters setting and training. The parameters of convolutional layers and full-connected layers are illustrated in Fig. 4. First, we train facial component classification network with initial learning rate of 0.02 and decayed 20% after every 8 epochs. The weight decay is 0.0001 and momentum is 0.9. Then we copy weights of shared convolutional module and initialize other layers with Gaussian initialization [45]. RPN generates candidate proposals with a mini-batch of 2 images, and region proposals are inputs of Fast R-CNN. We adopt 4-Step Alternating method proposed by [17] to train RPN jointly with Fast R-CNN, which begins with RPN. The initial learning rate setting is same as [17]. As for recurrent learning module, 256 hidden units are used in the LSTM layers and we set $T = 4$. Note that our model tests one facial image in 86 ms on a Nvidia GeForce GTX1080 Ti graphics card and Intel i7-4930 K CPU.

Evaluation metric. While testing, we evaluate the performance of eye landmarks detection of our model. Similar to the evaluation method of facial key points detection in [46], failure rate and RMSE (Root Mean Square Error) between ground truth coordinates of eye landmarks $S = (x_{g1}, y_{g1}, \ldots, x_{g7}, y_{g7})$ and the output eye relative shape

Table 2
MAP(%) results about each component by WSELD-sup and WSELD.

Categories	eyebrow	eye	nose	mouth
WSELD-sup	64.2	91.3	88.5	87.6
WSELD	36.7	72.5	41.4	38.9

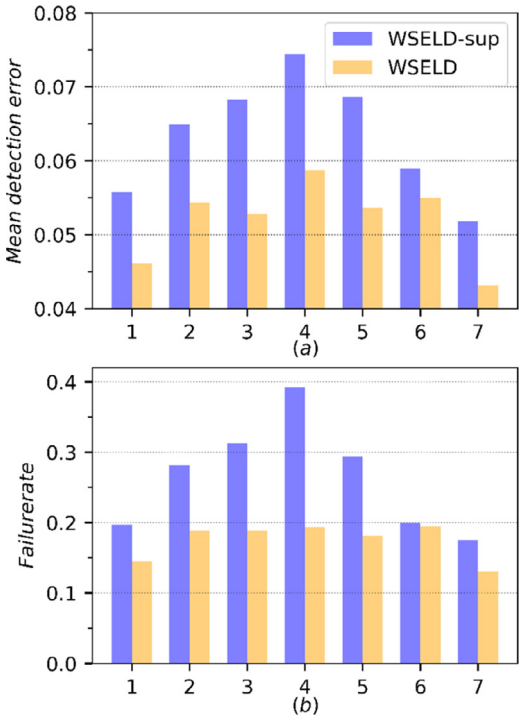

Fig. 6. Comparison of WSELD-sup and WSELD: mean detection error (a) and failure rate (b) about each eye point.

$\tilde{S} = (\tilde{x}_1, \tilde{y}_1, \ldots, \tilde{x}_7, \tilde{y}_7)$ are used for quantitative measurement. RMSE can be formulated as:

$$d_i = \sqrt{\left(x_{gi} - \tilde{x}_i\right)^2 + \left(y_{gi} - \tilde{y}_i\right)^2}/a$$
$$d_{RMSE} = \frac{1}{7}\sum_{i=1}^{7} d_i \tag{8}$$

where d_i indicates mean detection error about a point and a is normalization factor, which is the square root of the ground truth bounding box, computed as $a = \sqrt{g_w * g_h}$. Mean detection error about each point intuitively illustrates the promotion of our model by the submodules of faster R-CNN and recurrent learning module. Besides, samples with RMSE bigger than 0.08 are reported as failures.

4.1. Controlled experiments on our dataset

4.1.1. Experiments on faster R-CNN module

To evaluate the efficiency of faster R-CNN module, we create two models based on the same structure in Fig. 4(b) and both of them are without recurrent learning module. One model is trained by supervised and weakly supervised data, named as WSELD (Weakly Supervised Eye Landmarks Detection), and the other is completed only by supervised data, named as WSELD-sup. Both of them are tested on supervised testing set to validate the feasibility of weakly supervised eye landmarks detection algorithm.

First, mAP (mean Average Precision) about each component is recorded in Table 2. We find that two models are able to detect facial components, and mAP of eye bounding-box is optimal. We infer that our models pay more attention to eye landmarks detection through setting a little bigger hyperparameters for the branch of eye landmarks regression in loss function. Besides, WSELD outperforms WSELD-sup on detection accuracy of each facial component. We suppose that weakly supervised data also have the information of bounding-box for facial components, which are added into the training set of WSELD. While WSELD-sup is just trained on supervised data.

Table 3
Comparison of RMSE and Failure rate about eye shape of methods mentioned in experiments. The results are evaluated on the same test set.

Methods	RMSE	Failure rate
WSELD-sup	0.0632	0.2245
WSELD	0.0520	0.1415
T1	0.0453	0.066
T2	0.0420	0.0585
T3	0.0409	0.0565
T4	**0.0391**	**0.0405**

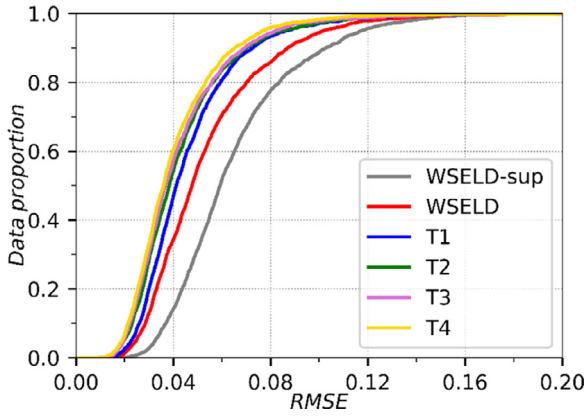

Fig. 7. Cumulative error curve with respect to RMSE for WSELD-sup, WSELD, WSELD-T1∼4, evaluated on the same test set. WSELD-T4 outperforms than other models.

However, we concern more about results of eye landmarks detection. We present mean detection error and failure rate about each point of WSELD-sup and WSELD in Fig. 6. It shows that two models can achieve eye landmarks detection using Faster R-CNN, and WSELD has lower mean detection error and failure rate about each point. Besides, mean detection error at point 7 is lower than the average of that about 2,3,5 and 6, which proves the necessity of our operation of calibrating the iris. Additionally, RMSE and failure rate are recorded in Table 3. RMSE of WSELD is 0.052, which significantly improved by 17.7% compared to that of WSELD-sup. Failure rate about the eye shape also has dropped a lot with the improvement of weakly supervised learning. These two factors clearly demonstrate that weakly supervised data is able to help our model extract better representations for eye landmarks detection, even without eye landmarks information in samples.

To further figure out the distribution of RMSE about test set, we give the cumulative error curve with respect to RMSE in Fig. 7. When RMSE is 0.08 and 0.04, the accuracy of detection is improved up to 9.3% and 58%, respectively. All results mentioned above demonstrate that the performance of WSELD is better than WSELD-sup. Qualitative results using the two models on the same tested facial image are provided in Fig. 8. Both WSELD and WSELD-sup can detect eye points in facial image, but location result of WSELD is much better than WSELD-sup. We see that detected points from WSELD are more accurately located at the eye

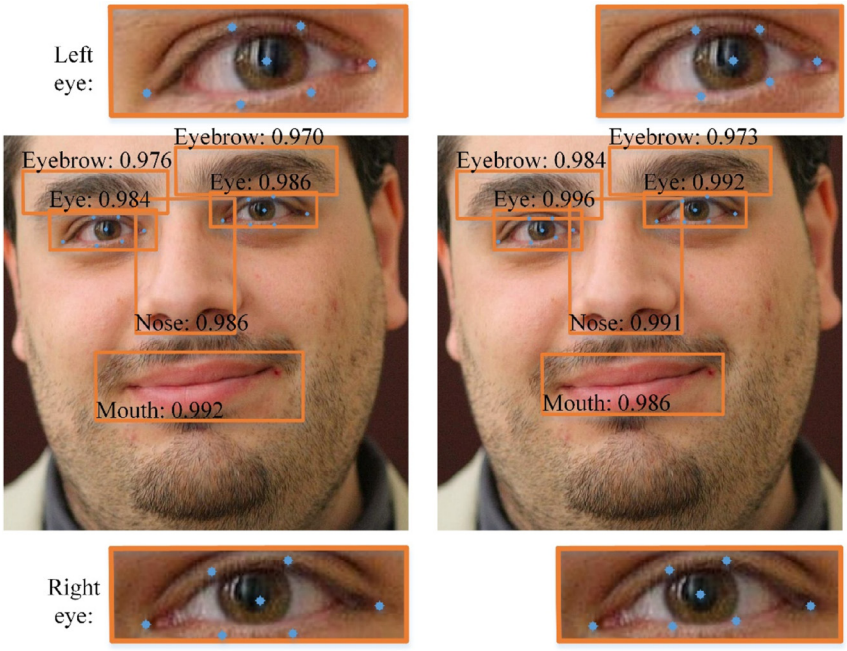

Fig. 8. Qualitative results of the two models. The left image is the detection result of WSELD-sup, while the right belongs to WSELD. The upper and lower figures are enlarged local patches of left and right eye, which can show the detection results in detail.

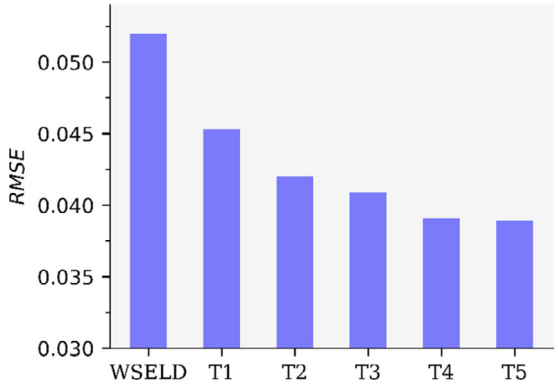

Fig. 9. Accuracy comparison of recurrent module with different time stage.

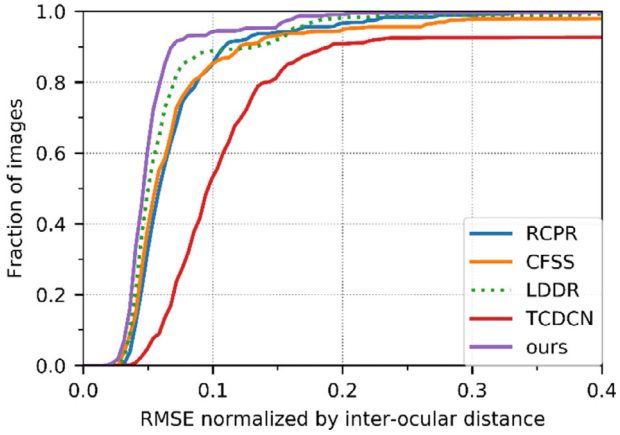

Fig. 10. Cumulative errors distribution curves with respect to RMSE normalized by inter-ocular distance on LFPW testing set.

contours, especially the point 2, 3, 5 and 6 getting closer to the intersection of eyelids and iris. While SEMI-sup performs not as well as the detected points are located around the eye contour.

4.1.2. Experiments on recurrent learning module

To examine the promotion of recurrent learning module for eye landmarks detection, we test the recurrent module with different time steps on testing set. Table 3 reports RMSE and failure rate of recurrent module with different time stage, and Fig. 7 shows the corresponding cumulative error distribution curve with respect to RMSE. We see that the RMSE of recurrent module with four time stages is 0.0391 and approximately improved by 24.8% compared to WSELD. As seen from Fig. 7, it is clear that recurrent module at the fourth stage also performs best, which proves that sharing middle level representation using LSTM layers can improve the accuracy of eye landmarks detection.

In Fig. 9, we report RMSE of recurrent module with different time stage. We find that RMSE decrements of recurrent module at the 1st (from WSELD to T1) and 2nd (from T1 to T2) are relatively larger than the latter stages. We infer that the information of middle level features is related to the shape increment. The earlier stage often deals with much different shapes while the output of later stage is close to the optimal shape. It can be further observed that more time stages in recurrent learning module gain a similar result as $T=4$. Therefore, we adopt four-time stages for our recurrent learning module by default.

4.2. Comparison between facial landmark detection methods

To examine effectiveness of the proposed algorithm for facial landmarks detection, we compare our method with the state-of-the-art methods on LFPW and 300 W testing set. We do not conduct model comparison on AFLW due to all samples that are regarded as weakly supervised data. Because the proposed method can only detect right and left eye landmarks in a single facial image, we convert eye landmarks error to 68 key points error by multiplying approximate coefficient. Table 4 records the comparative results of RMSE normalized by inter-ocular distance. It can be seen that the proposed method performs best and RMSE on LFPW

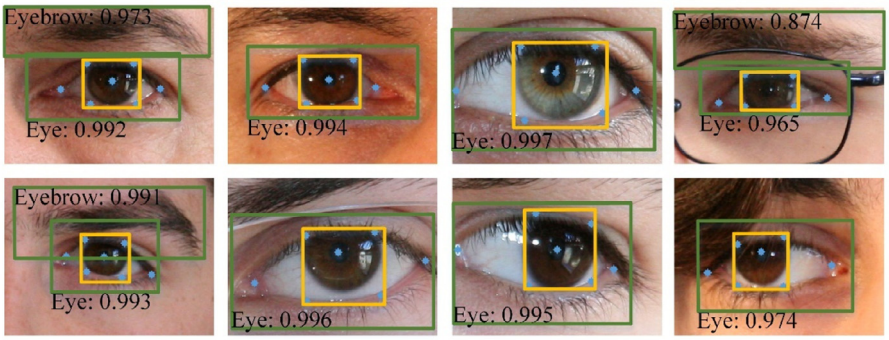

Fig. 11. Some detection results on UBIRIS.v2. Our algorithm could handle the scales and poses of iris and partial occlusion. The green bounding-box and the blue points are outputs of our algorithm. The yellow bounding-box indicates the poistion of the iris. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quantitative results of different models on LFPW testing set.

Methods	LFPW
RCPR [47]	6.56
SDM [48]	5.67
CFAN [49]	5.44
CFSS [50]	4.87
LDDR [51]	4.67
PCSR [52]	3.81
Ours	3.56

Table 5
Quantitative results of different models on 300 W testing set.

Methods	300-W (Full)	Methods	300 W (Full)
SDM [48]	7.52	RDR [25]	5.80
RCPR [47]	6.18	PCSR [52]	5.18
CFSS [50]	5.76	Seq-MT [53]	5.74
TCDCN [54]	5.54	FARN [56]	4.88
RAR [55]	4.94	SAN [28]	4.24
Thewlis et al. [29]	7.97	**Ours**	4.58

testing set is improved by relative 6.56% compared to PCSR [52]. This implies that our method has the ability to perform more robust and accurate by weakly supervised learning or refinement of recurrent learning module. As shown in Fig. 10, we compare the proposed method with RCPR [47], CFSS [50], LDDR [51], TCDCN [54] on LFPW testing set. It is obvious that our method outperforms the above mentioned state-of-the-art methods.

In order to compare performance of facial landmarks detection fairly on 300 W testing set, we get the bounding box provided by the official 300 W detector. Table 5 reports results of the proposed method compared to other facial landmarks detection methods on 300 W testing set. Our method performs comparably with many state-of-the-art methods. We find that the error of SAN [28] is lower than our method. This may due to the fact that SAN integrates Cycle-GAN to generate style-aggregated face images and then takes the original image and the style-aggregated one as two complementary inputs to train facial landmarks prediction module. Note that our method focuses on eye landmarks detection combined with object detection without supervisory information of key points provided by weakly supervised data. The object detection module of the proposed model can be replaced with more accurate methods.

Table 6
Results of iris segmentation for UBIRIS.v2.

Method	Segmentation error rate, E(%)
Hugo [57]	3.75
Tan et al. [58]	3.49
Mohammed et al. [59]	2.95
Tan and Kumar [60]	2.37
Ours	2.12

4.3. Comparison with state-of-the-art methods

In this section, we conduct two experiments on UBIRIS.v2 and MICHE databases to evaluate the performance of our method. The target of our model is to detect eye landmarks in facial images, which may not be directly compared with other methods on these iris databases. Therefore, we are going to make some conversions with the results of eye landmarks detection in the following experiments and give quantitative evaluations.

4.3.1. Experiments on UBIRIS.v2 database

UBIRIS.v2 contains 11,102 eye images captured in non-constrained conditions, with resolution (400×300) pixel. The images of this database vary in lighting conditions using natural and artificial lighting sources. To simulate noise factors, such as the scale and head orientations, volunteers were required to move slowly and to turn their heads to look at several marks. NICE.II is a subset of UBIRIS.v2, which includes 1000 eye images and the corresponding ground truth of segmentation maps to evaluate iris segmentation algorithms.

This dataset is employed to evaluation of iris segmentation algorithms, such as [57–60]. To compare the performance of iris localization, we should make post-processing schemes. First, the iris structure is located at the window with max region of predicted points 2, 3, 5 and 6 by using the proposed method. We show some randomly selected irises localized accurately using the proposed algorithm in Fig. 11. Our method can directly detect the eyes, even the eyebrows from local view of facial images. It demonstrates that our method can work on local view of facile images, which benefits from incorporating facial components detection with eye landmarks detection. Then based on this localization, we complete iris segmentation using the GrowCut technique proposed by [61]. We validate the ability of the proposed method in comparison with recent iris segmentation methods through measurement of segmentation error rate E [62]. Table 6 shows segmentation error rates of all methods. Note that results of other methods are derived from published papers. We observe that our method performs best with a segmentation error rate of 2.12%. The proposed method achieves at least improvement up to 11% over the accuracy of other iris

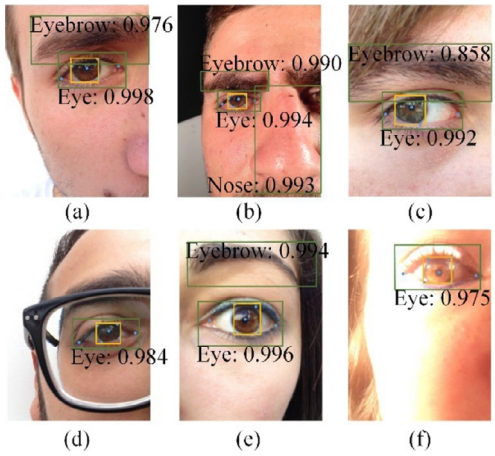

Fig. 12. Some randomly selected iris localization results from MICHE with challenging factors.

segmentation methods, which demonstrates that our method can outperform the state-of-the-art methods.

4.3.2. Experiments on MICHE database

In this setup, we test the proposed model on MICHE database, which is captured "in the wild" using mobile devices, including Samsung Galaxy Table 2, Samsung Galaxy S4 and Apple iPhone5. Eye images in MICHE are quite frontal face with the lower resolution. MICHE is more challenging due to several influenced noise factors like specular reflection, blur and occlusion of eyelids. However, there are not the ground truth of segmentation maps corresponding to the original iris images in MICHE database, so we provide qualitative results shown in Fig. 12. We find that our method is robust enough to handle different challenging factors, such as wide variations in lighting conditions (Fig. 12(f)), iris scales (Fig. 12(b)-(c)), background clutter including nose (Fig. 12(b)), two eyes (Fig. 12(b)), eyebrows (Fig. 12(a), (b), (c) and (e)) and eyeglasses (Fig. 12(d)). It clearly proves that the proposed method is effective for iris localization in a non-constrained imaging environment.

5. Conclusions

In this paper, we have proposed a weakly supervised eye landmarks detection algorithm. Our method is an attempt to apply object detection, which provide additional information for eye landmarks detection. It is significant that our method achieves predicting eye landmarks directly without prerequisites of facial landmarks localization. As compared to other methods through careful evaluation, our method provides acceptable performance of locating iris from eye regions under unconstrained environment.

It is worth mentioning that the proposed method is trained on facial images but is competent to detect eye region and landmarks simultaneously from severe occluded or local view of facial images when many existing face detectors fail to work. Our method mainly benefits from the special format data which contains two kinds of samples. One is the images with ground truth of bounding-box of facial components, corresponding classification labels and coordinates of eye landmarks while weakly supervised data do not have eye landmarks information. We believe that ideas for learning models with different types of training data can inspire other key points regression tasks [63,64].

However, some limitations for the proposed method need to be noticed. Our method combines object detection algorithm with eye landmarks detection task and recurrent learning module fine tunes eye landmarks based on the former prediction. This fact makes our method not competitive in terms of speed. Therefore, we would explore faster object detection algorithm to accelerate our method for our future work.

Declaration of Competing Interest

None

Acknowledgments

This work was funded by a project that partially funded by National Science Foundation of China (51675265) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The authors gratefully acknowledge this support.

References

[1] Q. Ji, Y. Xiaojie, Real time visual cues extraction for monitoring driver vigilance, in: International Conference on Computer Vision Systems, Berlin, Heidelberg, Springer Berlin Heidelberg, 2001, pp. 107–124.
[2] D. Young, H. Tunley, R. Samuels, Specialised hough transform and active contour methods for real-time eye tracking, 1995.
[3] K.M. Lam, H. Yan, Locating and extracting the eye in human face images, Pattern Recognit. 29 (5) (1996) 771–779.
[4] S. Kawato, N. Tetsutani, Real-time detection of between-the-eyes with a circle frequency filter, in: Proceedings of the 5th Asian Conference on Computer Vision (ACCV2002), 2, 2002, pp. 23–25.
[5] G.C. Feng, P.C. Yuen, Variance projection function and its application to eye detection for human face recognition, Pattern Recognit. Lett. 19 (9) (1998) 899–906.
[6] P.W. Hallinan, Recognizing human eyes, Spie. Proc. Geom. Methods Comput. Vis. 1570 (1991) 214–226.
[7] V. Laxmi, P.S. Rao, Eye detection using gabor filter and svm, in: 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 880–883.
[8] J. Huang, H. Wechsler, Eye detection using optimal wavelet packets and radial basis functions (rbfs), Int, J. Pattern Recognit. Artif. Intell. 13 (07) (1999) 1009–1025.
[9] P. Wang, M.B. Green, Q. Ji, J. Wayman, Automatic eye detection and its validation, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops, 2005, p. 164. -164.
[10] X. Xie, R. Sudhakar, H. Zhuang, On improving eye feature extraction using deformable templates, Pattern Recognit. 27 (6) (1994) 791–799.
[11] T. Ishikawa, S. Baker, I. Matthews, T. Kanade, Passive driver gaze tracking with active appearance models, in: Proceedings of the 11th World Congress on Intelligent Transportation Systems, 3, 2004, pp. 41–43.
[12] C.N. Meng, J.J. Bai, T.N. Zhang, S.J. Chang, Research on eye gaze estimation based on low-cost eye movement recorder, J. Optoelectron. Laser 24 (8) (2013) 1600–1605.
[13] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba, Eye tracking for everyone, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2176–2184.
[14] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4511–4520.
[15] B. Huang, R. Chen, Q. Zhou, X. Yu, Eye landmarks detection via two-level cascaded cnns with multi-task learning, Signal Process. 63 (2018) 63–71.
[16] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[17] S. Ren, K. He, R. Girshick, S. Jian, Faster r-cnn: towards real-time object detection with region proposal networks, International Conference on Neural Information Processing Systems, 2015.
[18] H. Proenca, S. Filipe, R. Santos, J. Oliveira, L. Alexandre, The ubiris.v2: a database of visible wavelength iris images captured on-the-move and at-a-distance, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1529–1535.
[19] M.D. Marsico, M. Nappi, D. Riccio, H. Wechsler, Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols, Pattern Recognit. Lett. 57 (2015) 17–23.
[20] U. Weidenbacher, G. Layher, P. Strauss, H. Neumann, A comprehensive head pose and gaze database, in: 2007 3rd IET International Conference on Intelligent Environments, 2007, pp. 455–458.
[21] B.A. Smith, Q. Yin, S.K. Feiner, S.K. Nayar, Gaze locking: passive eye contact detection for human-object interaction, Acm Symposium on User Interface Software & Technology, 2013.
[22] K.A. Funes Mora, F. Monay, J.-M. Odobez, Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras, in: Proceedings of the ACM Symposium on Eye Tracking Research and Applications, ACM, 2014.
[23] Q. Huang, A. Veeraraghavan, A. Sabharwal, Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets, arXiv:1508.01244.

[24] M.J.T. Reinders, R.W.C. Koch, J.J. Gerbrands, Locating facial features in image sequences using neural networks, in: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996, pp. 230–235.
[25] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, A. Kassim, Recurrent 3d-2d dual learning for large-pose facial landmark detection, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1633–1642.
[26] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1021–1030.
[27] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 379–388.
[28] A. Bulat, G. Tzimiropoulos, Super-fan: integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 109–117.
[29] J. Thewlis, H. Bilen, A. Vedaldi, Unsupervised learning of object landmarks by factorized spatial embeddings, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5916–5925.
[30] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, Y. Sheikh, Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 360–368.
[31] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: Proc. 31st International Conference on Machine Learning, 2014, pp. 1764–1772.
[32] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 3104–3112.
[33] G. Trigeorgis, P. Snape, M.A. Nicolaou, E. Antonakos, S. Zafeiriou, Mnemonic descent method: a recurrent process applied for end-to-end face alignment, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4177–4187.
[34] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, Robust facial landmark detection via recurrent attentive-refinement networks, in: European Conference on Computer Vision, Springer, 2016, pp. 57–72.
[35] A. Graves, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[36] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 545–552.
[37] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part II, Berlin, Heidelberg, Springer Berlin Heidelberg, 2012, pp. 679–692.
[38] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.
[39] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, A semi-automatic methodology for facial landmark annotation, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 896–903.
[40] M. Kstinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 2144–2151.
[41] L. van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. (2008) 2579–2605.
[42] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[43] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 00, 2014, pp. 580–587.
[44] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international Conference on Multimedia, ACM, 2014, pp. 675–678.
[46] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3476–3483.
[47] X.P. Burgos-Artizzu, P. Perona, P. Dollar, Robust face landmark estimation under occlusion, in: The IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1513–1520.
[48] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[49] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 1–16.
[50] S. Zhu, C. Li, C. Change Loy, X. Tang, Face alignment by coarse-to-fine shape searching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4998–5006.
[51] A. Kumar, R. Ranjan, V.M. Patel, R. Chellappa, Face alignment by local deep descriptor regression. arXiv:1601.07950.
[52] Q. Liu, J. Yang, J. Deng, K. Zhang, Robust facial landmark tracking via cascade regression, Pattern Recognit. 66 (2017) 53–62.
[53] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, J. Kautz, Improving landmark localization with semi-supervised learning, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1546–1555.
[54] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Learning deep representation for face alignment with auxiliary attributes, IEEE Trans. Pattern Anal. Mach. Intell. 38 (5) (2016) 918–930.
[55] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, Robust facial landmark detection via recurrent attentive-refinement networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision - ECCV 2016, Springer International Publishing, Cham, 2016, pp. 57–72.
[56] Q. Hou, J. Wang, R. Bai, S. Zhou, Y. Gong, Face alignment recurrent network, Pattern Recognit. 74 (2018) 448–458.
[57] H. Proenca, Iris recognition: on the segmentation of degraded images acquired in the visible wavelength, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1502–1516.
[58] T. Tan, Z. He, Z. Sun, Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition, Image Vis. Comput. 28 (2) (2010) 223–230.
[59] M.A.M. Abdullah, S.S. Dlay, W.L. Woo, J.A. Chambers, Robust iris segmentation method based on a new active contour force with a noncircular normalization, IEEE Trans. Syst. Man Cybern. 47 (12) (2017) 3128–3141.
[60] C. Tan, A. Kumar, Automated segmentation of iris images using visible wavelength face images, in: CVPR 2011 Workshops, 2011, pp. 9–14.
[61] V. Vezhnevets, V. Konouchine, Growcut: interactive multi-label nd image segmentation by cellular automata, Proc. Graph. 1 (2005) 150–156.
[62] H. Proenca, L.A. Alexandre, The nice.I: noisy iris challenge evaluation - part I, in: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–4.
[63] J. Gao, A.N. Evans, Expression robust 3d face landmarking using thresholded surface normals, Pattern Recognit. 78 (2018) 120–132.
[64] R. Ma, H. Hu, W. Wang, J. Xu, Z. Li, Photorealistic face completion with semantic parsing and face identity-preserving features, ACM Trans. Multimedia Comput. Commun. Appl. 15 (1) (2019) 28:1–28:18.

Bin Huang graduated from Nanjing University of Aeronautics & Astronautics (NUAA), Nanjing, China in 2015, and is currently working towards the Ph.D. degree in Measurement and Testing Technology & Instruments at State Key Laboratory of Mechanics & Control of Mechanical Structures, NUAA, Nanjing, China. His current field of interest focuses on machine learning, fatigue detection, and computer vision.

Renwen Chen graduated from Nanjing University of Aeronautics & Astronautics (NUAA), Nanjing, China in 1991 and received his Ph.D. degree from NUAA in 1999, both in Measurement and Testing Technology & Instruments. He also spent half a year as a visiting scholar in University of California, Berkeley, USA. His current research interests are in the field of machine learning, energy harvesting, wireless sensors networks and intelligent monitoring and control.

Qinbang Zhou graduated from Nanjing University of Aeronautics & Astronautics (NUAA), Nanjing, China in 2014, and is currently working towards the Ph.D. degree in Measurement and Testing Technology & Instruments at State Key Laboratory of Mechanics & Control of Mechanical Structures, NUAA, Nanjing, China. His current field of interest focuses on machine learning, fatigue detection, and computer vision. His research interests are machine learning, damage detection, and computer vision.

Wang Xu graduated from Nanjing University of Aeronautics & Astronautics (NUAA), Nanjing, China in 2016, and is currently working towards the Ph.D. degree in Measurement and Testing Technology & Instruments at State Key Laboratory of Mechanics & Control of Mechanical Structures, NUAA, Nanjing, China. His current field of interest focuses on machine learning, fatigue detection, and computer vision. His research interests are machine learning, super-resolution image reconstruction, and computer vision.