Learning to Predict Saliency on Face Images

Mai Xu*, Yun Ren, Zulin Wang
School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China
MaiXu@buaa.edu.cn

Abstract

This paper proposes a novel method, which learns to detect saliency of face images. To be more specific, we obtain a database of eye tracking over extensive face images, via conducting an eye tracking experiment. With analysis on eye tracking database, we verify that the fixations tend to cluster around facial features, when viewing images with large faces. For modeling attention on faces and facial features, the proposed method learns the Gaussian mixture model (GMM) distribution from the fixations of eye tracking data as the top-down features for saliency detection of face images. Then, in our method, the top-down features (i.e., face and facial features) upon the the learnt GMM are linearly combined with the conventional bottom-up features (i.e., color, intensity, and orientation), for saliency detection. In the linear combination, we argue that the weights corresponding to top-down feature channels depend on the face size in images, and the relationship between the weights and face size is thus investigated via learning from the training eye tracking data. Finally, experimental results show that our learning-based method is able to advance state-of-the-art saliency prediction for face images. The corresponding database and code are available online: www.ee.buaa.edu.cn/xumfiles/saliency_detection.html.

1. Introduction

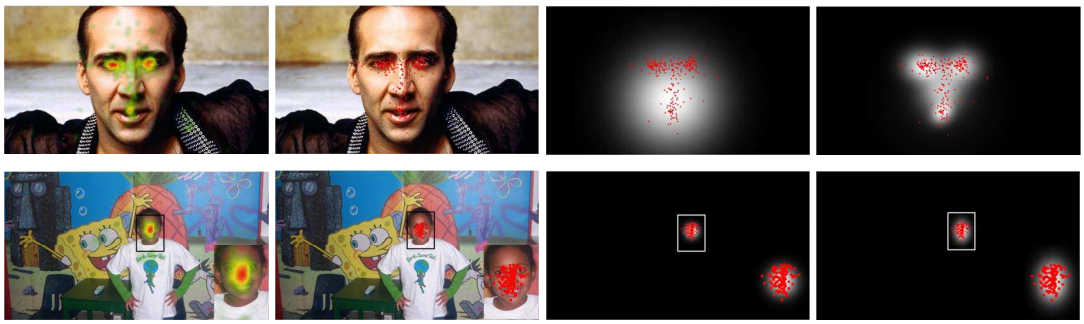
According to the study on the human visual system (HVS) [17], when a person looks at a scene, she/he may pay much visual attention to a small region (the fovea) around a point of eye fixation with high resolutions. The other regions, namely the peripheral regions, are captured with little attention at low resolutions, such that humans can survive from the processing of tremendous visual data. Visual attention therefore is a key to perceive the world around humans, and it has been widely studied in psychophysics, neurophysiology, and even computer vision societies [1]. With computation on features of either images or videos, saliency detection is an effective way to predict the human visual attention attracted by different regions of a scene. As the output of saliency detection, the saliency map of an image or a video frame has been widely applied in object detection [3], object recognition [8], image retargeting [20], image quality assessment [6], and also image/video compression [25].

The existing methods on saliency detection can be classified into two categories: bottom-up and top-down methods. The representative bottom-up method on detecting image saliency is Itti's model [13], which combines center-surround features of color, intensity, and orientation together. Afterwards, Koch and Ullman [24] extended Itti's model by incorporating the proto-object inference in the saliency map. Most recently, there has been extensive advanced work (e.g., [5, 7, 10, 11, 26]) on bottom-up saliency detection.

In fact, top-down visual features play a crucial role in determining the saliency of a scene. Hence, the top-down saliency detection methods have been broadly studied in [4, 22, 28]. Cerf *et al.* [4] found out that face is an important top-down feature to attract visual attention, as in their experiments faces were fixed on in 88.9% within first two fixations (7 subjects viewing 150 face images). Therefore, they proposed to combine Viola & Jones (VJ) face detector [23] with Itti's model [13] for improving the saliency detection accuracy over face images. Since it is more reasonable to learn how important face is for attracting visual attention, several state-of-the-art methods [12, 15, 27] have been proposed to apply machine learning algorithms in top-down saliency detection of Cerf's work [4]. For example, Zhao [27] utilized the fixations on face images to quantify the weight of the face channel on attracting visual attention. Most recently, Jiang *et al.* [14] has extended Cerf's work [4] to saliency detection in a scene with multiple faces, i.e., saliency detection in a crowd. In their work, multiple kernel learning (MKL) is applied to learn a more robust discrimination between salient and non-salient regions in multi-face scenes, for detecting saliency in a crowd.

Although the existing work has taken into account one or more faces on saliency detection, it does not explore the distribution of eye fixations within faces. As shown in Fig-

*Mai Xu and Yun Ren contribute equality to this work.



(a) Fixation Heatmap (b) Fixations on face (c) Isotropic GM (d) Learnt GMM

Figure 1. Examples for saliency prediction vs fixations in face region, selected from [15]. The red dots represent the fixations recorded by the eye tracker. Note that both saliency and fixations belonging to face regions are displayed.

ure 1, a simple isotropic Gaussian model (GM) assumption for saliency distribution in face [4, 27] has the limitation on modeling visual attention attracted by faces. As can be seen in this figure, for images with small faces, non-isotropic GM is more accurate in modeling saliency distribution inside face. For images with large faces, a single GM is not effective, as the fixations tend to cluster around the facial features (e.g., eyes). Accordingly, saliency distribution, in the form of Gaussian mixture model (GMM), need to be learnt from eye fixations on face images. Figure 1-(d) shows that the saliency with the learnt GMM distribution is more consistent with the ground truth visual attention. Specifically, one non-isotropic Gaussian component should be utilized for images with small faces, whereas more than one components can be applied for images with large faces. This paper thereby proposes a learning-based saliency detection method, which learns various GMMs and the corresponding weights across different face sizes[1], for predicting visual attention on free-viewing face images.

The main contributions of this paper are listed as follows:

- We establish a large eye tracking database for visual attention analysis on face images, in which 510 images with faces at different sizes were free-viewed by 24 subjects. The ground truth fixations on viewing all 510 images are available[2]. The analytical results on our database reveal that humans tend to be attracted by faces. Specifically, when the face sizes are large, the majority of visual attention on faces is drawn by facial features. Such results motivate our learning-based method on saliency detection of face images.
- We model human visual attention attended to face regions using GMM distribution, which is learnt from eye fixations of training images in our database. Note that the visual attention model of this paper is for free-view scenario without any specific task. Specifically, we utilize Expectation Maximization (EM) algorithm [18] to learn GMM distribution of saliency in face region from the ground truth fixations. Based on the learnt GMM, two feature channels (on face and facial features) are integrated as the top-down information in saliency detection. For the integration, we argue that weights of the proposed top-down feature channels depend on face size, and they can also be learnt from the training face images.

[1] In this paper, face size means the proportion of pixel number of the face region to that of whole image.

[2] $www.ee.buaa.edu.cn/xumfiles/saliency\_detection.html$.

2. Database and analysis

Face, as the top-down cue [4], is of great importance to draw visual attention over face images. It is further intuitive that the facial features, such as eyes, may attract a large amount of visual attention. Thus, this section concentrates on figuring out how significant the face and facial features are to attract visual attention. Section 2.1 discusses the eye tracking database we established for the statistical analysis. In Section 2.2, a method on automatically extracting the face and facial features is presented, as the preliminary for our statistical analysis of visual attention. Section 2.3 analyzes the importance of face and facial features to visual attention, via investigating the data of our eye tracking database.

2.1. Database of eye tracking on face images

For the analysis of visual attention on face images, we conducted the eye tracking experiment to establish a database of eye tracking on various face images. In our database, 510 face images were randomly selected from Google with the following criteria. (1) The original resolution of all images is 1920×1080. (2) All images contain only one frontal face, in which the turning degree of head is less than $45°$. (3) The sizes of faces in 510 images vary from 0.0016 to 0.3018. Figure 2 shows the various sizes of faces across those 510 images. Note that the images in Figure 2 are sorted in accordance with the ascending order of their face sizes.

There were a total of 24 subjects (14 male and 10 female, aging from 19 to 35) served as observers in the eye tracking experiment. All subjects have either corrected or uncorrected normal eyesight. Note that two among 24 subjects were experts, who worked on the research field of saliency detection. The other 22 subjects did not have any background in saliency detection, and they were native to the purpose of the eye tracking experiment.

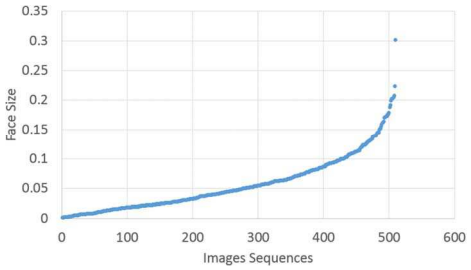


Figure 2. The distribution of face sizes in all 510 images, where the images are sorted by increased face sizes.

In the eye tracking experiment, a Tobii TX300 eye tracker, integrated with a monitor of 23-inch LCD displaying screen, was used to record the eye movement at a sample rate of 300 Hz. The resolution of the monitor was set to be 1920×1080, the same as the resolution of images. All subjects were seated on an adjustable chair at a distance of 60 cm from the monitor of the eye tracker. Therefore, the visual angle of the stimuli was about $26.8° \times 46.0°$. Before the experiment, subjects were instructed to perform the 9-point calibration for the eye tracker. During the experiment, each image was presented for 4 seconds, followed by a 2-second black image for a drift correction. All subjects were asked to free-view each face image. To avoid eye fatigue, the images were equally divided into 3 groups, each of which contained 170 images. After viewing one group of images, subjects had a 5 minute rest, and then were required to recalibrate the eye tracker before viewing the next group of images. Note that the displaying orders of both groups and images were random to further reduce the influence of eye fatigue on eye tracking results.

After the experiment, 151,511 fixations were collected. Averagely, each image had about 300 fixations. All the images, eye tracking data, and corresponding Matlab code are available on the Web to provide the ground truth data for saliency detection research.

2.2. Automatic detection on face and facial features

For analyzing the eye fixations on different parts of face, the regions for face and facial features have to be extracted in a face image. Generally speaking, our extraction technique is based on a real-time face alignment method [21]. To be more specific, several key feature points obeying the point distribution model (PDM) are located in an image using the method in [21], which combines the local detection (texture information) and global optimization (facial structure) together. Here, 66 key feature points, produced by the PDM, are connected to precisely identify the contours and regions of face and facial features.

2.3. Analysis of visual attention on face and facial features

Now, we move to analyzing visual attention on face and facial features, based on the statistics of our eye tracking database. Note that all 510 images with 151,511 fixations are used for the statistical analysis. In order to quantify visual attention on face, we plot in Figure 3 the percentages of fixations over all 510 images falling into face and background, respectively. We also plot in Figure 3 the proportions of pixels belonging to face and background, respectively. Note that faces were extracted using the method mentioned above. From this figure, we can see that although faces averagely take up 5.7% of whole images, they attract 62.3% of eye fixations. This verifies that the visual attention on face is significantly more than that on background.

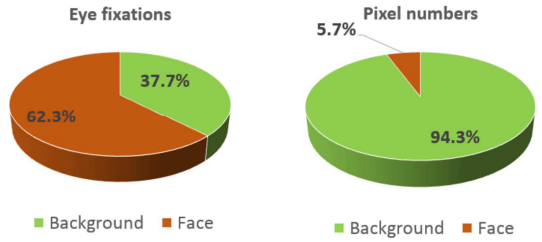


Figure 3. Proportions of eye fixations and pixel numbers for the regions of face and background.

Beyond, there is an insight that visual attention on face increases along with the enlarged face size in the image. To validate such an insight, we show in Figure 4 the proportions of fixations on faces versus face sizes, for all 510 images in our database. As can be seen from this figure, all points for proportion of fixations on face are above the random hit curve. Here, the random hit curve means the probability that a fixation randomly falls into the region of face. Again, this implies that face is with rather large saliency in an image. Besides, one may see from Figure 4 that the increase of fixation fitting curve is much faster than that of random hit, alongside the enlarged face size. Therefore, it can be concluded that much more attention is paid to face once the face is viewed at a large size.

Next, we discuss the statistical analysis on the eye fixations falling into different regions of face, to investigate the visual saliency of facial features, i.e., left eye, right eye, nose, and mouth. It is intuitive that the facial features are of great significance to visual attention, when the image is displayed with a close up view of face. Thus, it is interesting to find out the fixation proportions of facial features at various face sizes. Figure 5 shows the proportions of fixations on each facial feature versus face sizes, over all 510 images. From this figure, we can find out that more attention is drawn in all facial features than the random hit. Besides, it can be also observed that the fixation fitting curves for facial features, especially eyes, increase more sharply than the random hit, when face approaches large size. However, there is no proportion growth of fixations on nose region. It is probably because the visual attention shifts from face center (i.e., nose) to other facial features, such as eyes. In general, we can draw the conclusion that facial features of eyes and mouth are more salient, when the face has a large size in the image.

Together, Figures 3, 4, and 5 suggest that face and facial features have potential on drawing the majority of at-

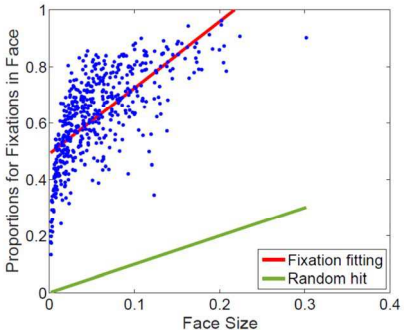


Figure 4. Statistical results on fixations belonging to face at different sizes for all 510 images. Note that the values of vertical and horizontal axes are proportions of fixations and size belonging to face within an image. Here, each point stands for the proportion of fixations belonging to face at one image. Then, the red line is the linear fitting curve on those points. Besides, the green line of random hit indicates the proportion for fixations randomly falling into the face region. Obviously, it is the same as the proportion of face region to the whole image.

tention, and that the visual attention on face and facial features (except the nose) rises along with the enlarged face size. Therefore, both face and facial features need to be taken into consideration for saliency detection, and the weights corresponding to these two channels should be relevant to face sizes. In the next section, the proposed method is to be introduced, which adds the channels of face and facial features to the conventional Itti's model [13].

3. The proposed method

This section mainly works on the proposed method for modeling saliency on face and facial features. In Section 3.1, we discuss preprocessing on the fixations for learning GMM. Next, GMM is learnt from the preprocessed training fixations, to be discussed in Section 3.2. Then, we present in Section 3.3 the saliency detection method based on the learnt GMM. Finally, in Section 3.4 we propose the way of obtaining optimal weights learnt from our database.

3.1. Preprocessing

For learning GMM, preprocessing has to be conducted to calibrate and normalize the eye fixations. Specifically, to avoid the uncertainty of face positions in different images, all fixations belonging to face region have to be calibrated in the following way.

As seen from Figure 6, Point A, the upper left point of PDM, is set to be the original point of the fixation coordinate in the face. Then, the coordinates (x, y) of fixations are calibrated to be (x^*, y^*) via translation:

$$\begin{cases} x^* = x - x_A \\ y^* = y - y_A, \end{cases} \tag{1}$$

where (x_A, y_A) is the coordinate for Point A.

Next, to deal with varying sizes of faces and facial features, fixations need to be normalized based on the width of face. To be more specific, the Euclidean distance l between Points A and B (as shown in Figure 6) is calculated

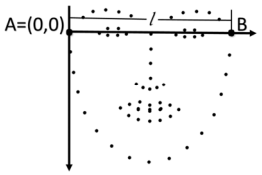


Figure 6. Coordinate calibration and normalization on 66-point PDM.

to be the face width, as the unit length for fixation coordinates. As such, the normalized coordinates (x', y') can be calculated as follows,

$$\begin{cases} x' = \dfrac{x^*}{l} \\ y' = \dfrac{y^*}{l}. \end{cases} \tag{2}$$

Finally, the positions for eye fixations attended to faces can be represented in a uniformed coordinate system. This way, all fixations in faces from different images can be processed together for learning GMM.

3.2. Learning GMM

As aforementioned, the facial features attract a large amount of visual attention, once the face is of large size. Therefore, we can use the GMM to model the facial feature channel, which has large-valued saliency within facial features. Assuming that $\mathbf{x} = (x', y')$ is the calibrated and normalized coordinate of point (x, y) within a face, the GMM can be written as a linear superposition of Gaussian components in the form:

$$\sum_{k=1}^{K} \pi_k \mathcal{N}_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3}$$

and

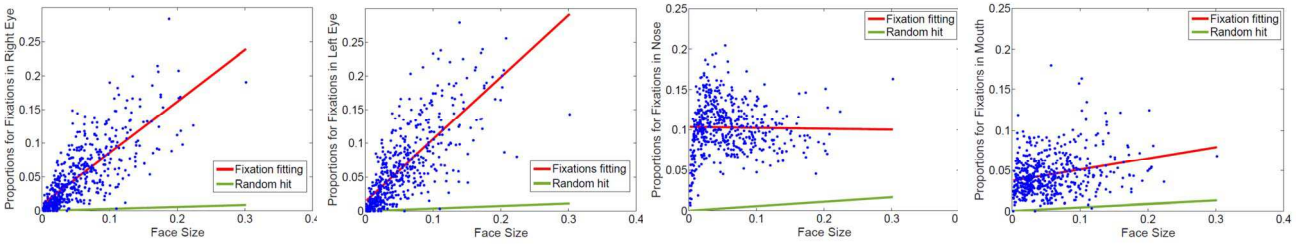
$$\mathcal{N}_k(\mathbf{x}) = \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\}, \tag{4}$$

where π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are the mixing proportion, mean, and variance of the k-th Gaussian component. In (3), K is the total number of Gaussian components.

In fact, the GMM can be learnt from fixations of eye tracking data. Here, the EM algorithm [18] is applied to learn the GMM on the calibrated and normalized fixations falling into face regions. For the face channel, the similar way is utilized to learn GMM distribution of face, where only one Gaussian component, corresponding to face, is considered. For the learnt results of GMMs on both face and facial feature channels, refer to Section 4.

3.3. Saliency detection

Given the learnt GMM, the top-down conspicuity maps on face channel ($\mathbf{F}$) and facial feature channel ($\mathbf{G}$), denoted by $\mathcal{C}(\mathbf{F})$ and $\mathcal{C}(\mathbf{G})$, can be worked out on the basis of (3) and (4). However, for saliency detection the mean values $\boldsymbol{\mu_k}$ in (3) and (4) are replaced by the central points of



(a) Right eye (b) Left eye (c) Nose (d) Mouth

Figure 5. Fixations versus size of each facial feature for all 510 images. Note that the values of vertical axis are portions of fixations falling into each facial feature, whereas the values of horizontal axis stand for portions of face size in an image. Here, each point means the proportion of fixations belonging to the corresponding facial feature at one image. Then, the red line is the linear fitting curve on those points. Besides, the green line of random hit indicates that proportion for fixations randomly fall into the facial feature region, such that it is the same as the proportion of facial feature size to the whole image size.

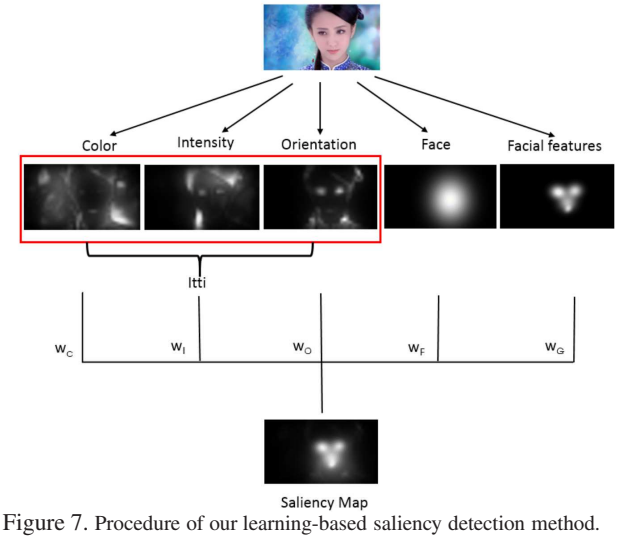


Figure 7. Procedure of our learning-based saliency detection method.

facial features, when the number of Gaussian components is 4. This is because there may exist the deviation between the statistical centroids of Gaussian components and the detected central points of facial features (i.e., eyes, nose, and mouth). Note that the face detection method is mentioned in Section 2.2.

Next, similar to [4], the top-down conspicuity maps are integrated with the bottom-up conspicuity maps of color ($\mathbf{C}$), intensity ($\mathbf{I}$), and orientation ($\mathbf{O}$). As a result, the final saliency map $\mathbf{M}$ can be generated by

$$\mathbf{M} = w_C\mathcal{C}(\mathbf{C}) + w_I\mathcal{C}(\mathbf{I}) + w_O\mathcal{C}(\mathbf{O}) + w_F\mathcal{C}(\mathbf{F}) + w_G\mathcal{C}(\mathbf{G}), \tag{5}$$

where $\mathcal{C}(\cdot)$ is the normalized conspicuity map on each feature channel. $\mathcal{C}(\mathbf{C})$, $\mathcal{C}(\mathbf{I})$, and $\mathcal{C}(\mathbf{O})$ can be obtained by the method in [24], whereas $\mathcal{C}(\mathbf{F})$ and $\mathcal{C}(\mathbf{G})$ need to be yielded upon the learnt GMM as aforementioned. In addition, $\mathbf{w} = [w_C, w_I, w_O, w_F, w_G]^T$ are weights corresponding to feature channels. They can be computed by least square fitting. For more details on computing these weights, refer to the next subsection. Figure 7 shows an example of overall procedure on our learning-based saliency detection method.

3.4. Learning optimal weights

Now, the remaining task for saliency detection with (5) is to determine weights $\mathbf{w} = [w_C, w_I, w_O, w_F, w_G]^T$ for each conspicuity map. In this subsection, we focus on the computation on learning optimal weights $\mathbf{w}$ from the training data of our eye tracking database. Let $\mathbf{m}_h$ be the vectorized human fixation map of a training image. Given $\mathbf{m}_h$, we follow the way of [27] to obtain weights $\mathbf{w}$ for each training image, by solving the following ℓ_2-norm optimization formulation:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{V}\mathbf{w} - \mathbf{m}_h\|_2, \quad s.t. \quad ||\mathbf{w}||_1 = 1, \mathbf{w} \geq 0, \tag{6}$$

where $\mathbf{V}$ is a matrix with each column denoting the vectorized conspicuity maps of $\mathbf{C}$, $\mathbf{I}$, $\mathbf{O}$ $\mathbf{F}$, and $\mathbf{G}$. Note that for each single image, (6) is solved to obtain an optimal weight $\mathbf{w}$ corresponding to this image. To solve (6), the disciplined convex programming approach [9] is utilized in our method. Then, the optimal weights can be obtained for each single training image. Note that the weight optimization in our method is different from that of [27] which works on the weights by fitting all training images.

Next, given the learnt weights for each individual image, we concentrates on working out the optimal weights of saliency detection, in light of different weights $\mathbf{w}$ of various trailing images. Specifically, we find that the optimal weights are dependent on face sizes. This is also consistent with the observation of Section 2.3, in which both face and facial features tend to attract much more attention when face is with large size. Thereby, it is worth figuring out the relationships between w_F and face size, and between w_G and face size. Here, the polynomial fitting is applied to model such relationship. Consequently, assuming that s is the face size, w_F and w_G can be expressed as follows,

$$w_F(s) = \sum_{i=0}^{I} a_i s^i, \tag{7}$$

and

$$w_G(s) = \sum_{i=0}^{I} b_i s^i, \tag{8}$$

where $\{a_i\}_{i=1}^I$ and $\{b_i\}_{i=1}^I$ are the parameters of quadratic functions to fit for w_F and w_G, respectively. As analyzed in Section 4, $I = 4$ is capable of producing the precise fitting on the pairs of weight and face size. Therefore, the

fourth order polynomial fitting is applied in this paper, and the values for $\{a_i\}_{i=1}^4$ and $\{b_i\}_{i=1}^4$ are to be discussed in Section 4.

After achieving w_F and w_G, other weights w_C, w_I, and w_O are averaged over all training images to acquire the ratios between them. Then, once w_F and w_G have been calculated by (7) and (8), w_C, w_I, and w_O can be determined according to the averaged ratios, with the constraint on $||\mathbf{w}||_1 = 1$. Values for the learnt parameters and ratios to yield weights $\mathbf{w}$ are to be reported in Section 4. Finally, the saliency map of a face image can be worked out via (5) with the learnt optimal weights.

4. Experimental results

In this section, experimental results are presented to evaluate the saliency detection performance of our method. In Section 4.1, we provide the training results on the GMMs and weights, which were learnt from ground truth fixations. In Section 4.2, we show the testing results of our method, in comparison with other 8 state-of-the-art methods: Itti *et al.* [13], Cerf *et al.* [4], Zhao *et al.* [27], Judd *et al.* [15], Duan *et al.* [5], Hou *et al.* [11], Erdem *et al.* [7], and Zhang *et al.* [26]. In the experiments, the area under ROC curve (AUC), normalized scanpath saliency (NSS) [19], and linear correlation coefficient (CC) [1] on all test images, were compared for evaluating the accuracy of saliency detection. In addition, the saliency maps of several test images are also provided for the comparison.

4.1. Training Result

In our experiment, we divided our eye tracking database of 510 images (as presented in Section 2.1) into training and test sets. For the training set, 360 images with 106,067 fixations were selected. For the test set, the remaining 150 images were chosen, which have 45,444 fixations. Note that there is no overlap between the training and test sets.

Learnt GMMs. In our experiments, we used the method of Section 3.2 to learn the GMMs for both face and facial feature channels of saliency detection, from the ground truth fixations of all 360 training images. For the face channel, the GMM was learnt with only one Gaussian component. The mean of the Gaussian component is simply assumed to be the position of nose tip point in each image (detected by the face alignment method [21]), as it can be seen as the center of face. Then, the covariance matrix for the Gaussian component was learnt from training data, and its learnt values are

$$\Sigma_1 = \begin{pmatrix} 0.024 & 0 \\ 0 & 0.039 \end{pmatrix}. \quad (9)$$

As can be seen above, there exists the anisotropy in learnt GMMs, rather than the assumption on isotropy of Gaussian distribution in [4].

For the facial feature channel, the number of Gaussian components has to be confirmed first. To determine the number of Gaussian components, we plot in Figure 8 the distributions of the learnt GMMs, with different numbers of Gaussian components. From this figure, we can see that the contours for GMMs with more than three components are similar. Accordingly, four-component GMM is utilized in our saliency detection method. This is also consistent with our analysis in Section 2.3 that visual attention tends to cluster around four facial features (i.e., left and right eyes, nose, and mouth). Hence, we assume that means of Gaussian components are the positions of the centers of facial features. The parameters of the learnt GMM in our learning-based method are tabulated in Table 1.

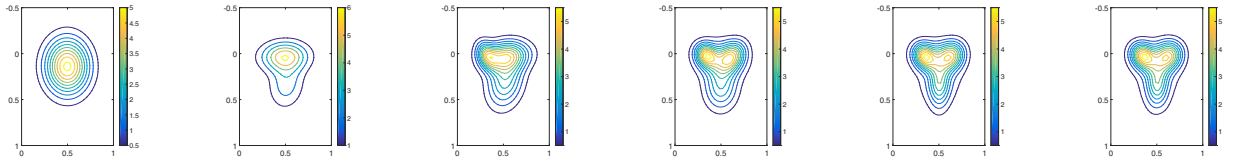
Learnt weights. Next, we obtained the optimal weight of each channel for the conspicuity maps of each individual image, using the optimization method of Section 3.4. As aforementioned, the optimal weights w_F and w_G for face and facial feature channels depend on the face size. Figures 9-(a) and -(b) plot the pairs of the face size and the corresponding optimal weight. Also, the curves on fitting those pairs of weight and face size are shown in Figures 9-(a) and -(b). We further show in Figure 9-(c) the Pearson's correlation coefficient (PCC) [16] on evaluation the fitting performance. It can be seen from this figure that PCC is nearly convergent for both face and facial feature channels, once the the order of polynomial fitting is greater than 3. In our experiments, the fourth order polynomial fitting were therefore adopted. Then, the values for fitting coefficients a_5, a_4, a_3, a_2, a_1 and a_0 of (7) are 6345.8, -2931.2, 491.0, -36.4, and 1.1, and values for b_5, b_4, b_3, b_2, b_1 and b_0 of (8) are -6474.3, 3146.4, -545.1, 38.6, and -0.1. Beyond, the ratio for $w_C : w_I : w_O$ is $8 : 3 : 30$, as the averaged optimal weights of color, intensity, and orientation channels are 0.016, 0.006, and 0.06. Finally, the saliency maps of all test images can be worked out by (5), with the aforementioned GMMs and optimal weights.

4.2. Testing Results

AUC. In order to quantify the accuracy of saliency detection, we tabulate in Table 2 the AUC results of our and other 8 methods. In this table, the AUC values are averaged over all 150 test images. Here, the results with and without center bias (CB) are provided. For fair comparison, all methods used the same CB filter [5]. As seen from Table 2, the methods with top-down features, i.e., Cerf *et al.* [4], Judd *et al.* [15], Zhao *et al.* [27] and ours, perform better than the bottom-up methods. This is because face, as a high-level feature, is crucial for improving saliency detection accuracy. Furthermore, our method outperforms other 8 state-of-the-art top-down and bottom-up methods in terms of AUC. Especially, there is 0.02 AUC improvement over Zhao *et al.* [27], which also integrates the the top-down face chan-

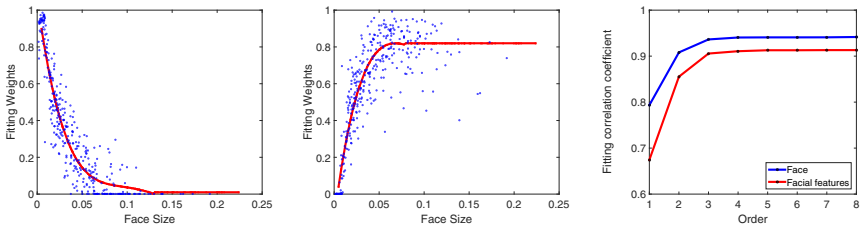
Table 1. The parameters of the learnt GMM

	k=1	k=2	k=3	k=4
features	right eye	left eye	nose	mouth
π_k	0.192	0.306	0.222	0.280
Σ_k	$\begin{pmatrix} 0.007 & 0.001 \\ 0.001 & 0.009 \end{pmatrix}$	$\begin{pmatrix} 0.013 & -0.002 \\ -0.002 & 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.035 & 0.003 \\ 0.003 & 0.032 \end{pmatrix}$	$\begin{pmatrix} 0.011 & -0.001 \\ -0.001 & 0.033 \end{pmatrix}$



(a) 1 component (b) 2 components (c) 3 components (d) 4 components (e) 5 components (f) 6 components

Figure 8. Contours of GMMs with various numbers of Gaussian components, learnt in our experiments.



(a) Face channel (b) Facial feature channel (c) PCC

Figure 9. Fitting on pairs of weight and face size. In (a) and (b), the blue dots stand for each pair of the optimal weight and its corresponding face size for all 360 test images, and the red lines are the fourth order polynomial curves on fitting all the blue dots. In (c), the orders of polynomial fitting curves versus PCC of fitting are plotted.

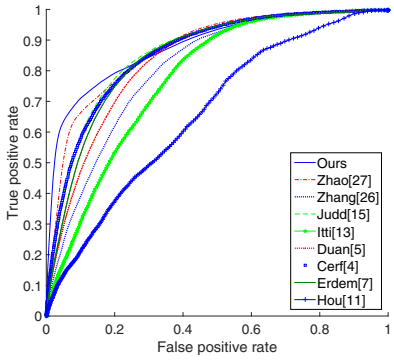


Figure 10. Average ROC curves for all 150 test face images, by our and other state-of-the-art methods. Note that all methods here are without any CB.

nel and learns its corresponding weight from training data. The possible reason for our method outperforming Zhao *et al.* [27] is that (1) the GMM distribution of saliency of face region is learnt from training data and then incorporated in our method, and that (2) the weights of top-down channels are learnt regarding face size. Moreover, we show in Figure 10 the ROC curves of saliency detection by our and other 8 state-of-the-art methods. Clearly, our method is superior to other methods.

NSS and CC. For a more comprehensive evaluation [2], we move to the comparison of NSS and CC metrics for saliency detection on all test images. NSS is computed to imply the relevance between fixation locations and saliency predictions, and CC measures the strength of a linear relationship between human fixation map and saliency map. The averaged NSS and CC results (with their standard deviations) of saliency detection by our and other state-of-the-art methods are also tabulated in Table 2. Note that methods with a larger NSS value or a CC value closer to +1/-1, can better predict the human fixations. Therefore, it can be seen from this table that our method performs significantly better than other state-of-the-art methods, in terms of both NSS and CC metrics. Specifically, there are at least 1.02 improvement of NSS and 0.17 enhancement of CC in our method without CB modeling. For saliency detection with CB modeling, similar NSS can CC improvement can be found in our method.

Saliency map. Figure 11 shows the saliency maps of 10 randomly selected test images, detected by eye tracking data, our, and other 8 methods. From this figure, we can see that compared to all other methods, our method is able to well locate the saliency regions, much closer to the maps of human fixations. To be more specific, for images with small face (i.e., from first to fourth rows), the saliency maps by our method are much more similar to those of human fixations than other methods, as the learnt non-isotropic Gaussian distribution of saliency in face region is adopted. For images with large face (i.e., from fifth to tenth rows), our method yields the appropriate maps, which well reflect the saliency distribution of regions of face and facial features using the learnt GMM. Moreover, our method is adaptive to predict human attention on faces with different sizes, since the optimal wights for face and facial feature channels in our method can be adjusted according to face size.

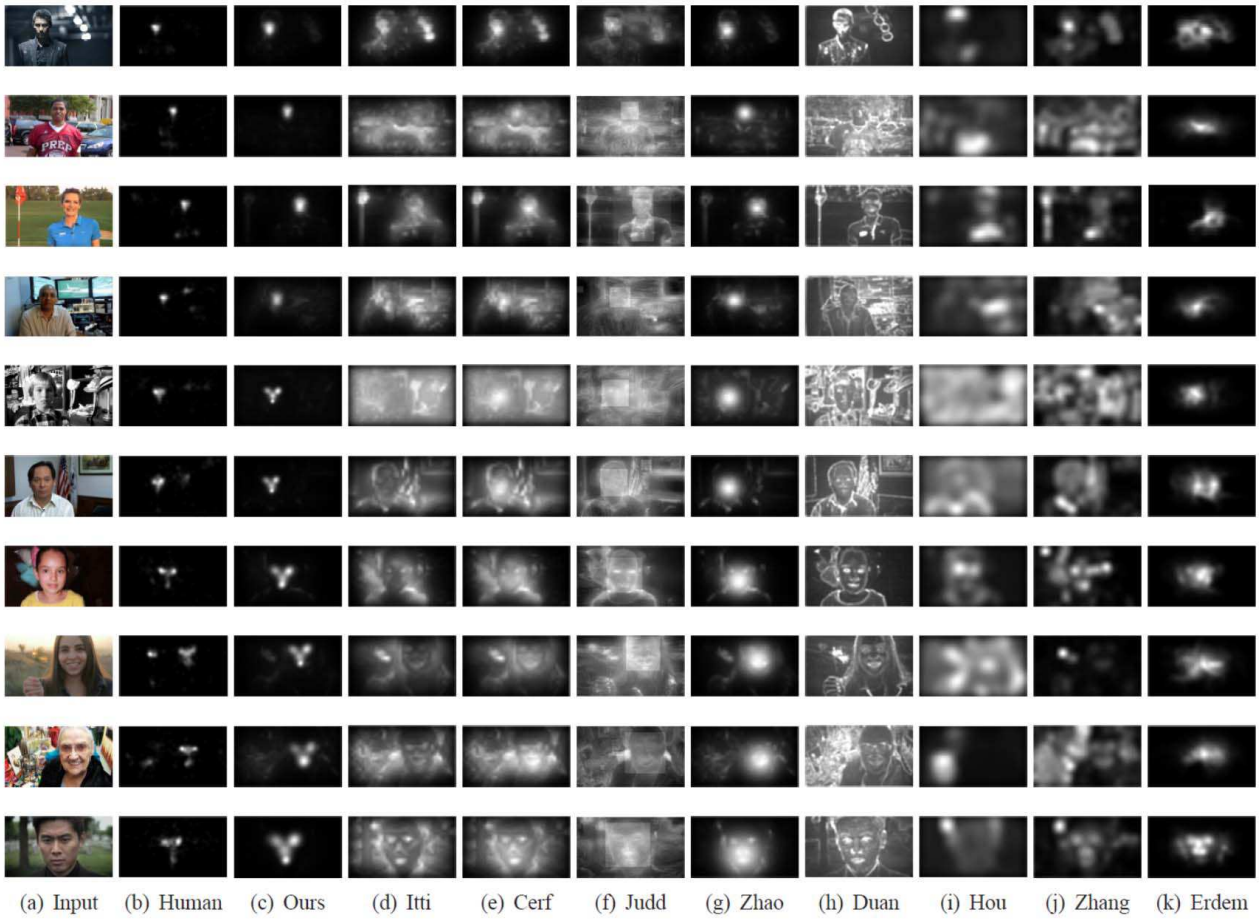


Figure 11. Saliency maps of several face images, produced by our and other state-of-the-art methods as well as by human fixations. Note that these images (from top to bottom) are sorted in the ascending order of face sizes. Also, note that the saliency maps here are without any CB.

Table 2. The comparison of our and other methods with and without CB model, for mean values (standard deviation) of AUC, NSS, and CC

Metrics	CB Model	Our method	Itti [13]	Cerf [4]	Judd [15]	Zhao [27]	Duan [5]	Hou [11]	Erdem [7]	Zhang [26]
AUC	Without CB	**0.90(0.04)**	0.78(0.10)	0.86(0.06)	0.77(0.10)	0.88(0.05)	0.79(0.09)	0.70(0.16)	0.84(0.06)	0.82(0.11)
	With CB	**0.90(0.04)**	0.82(0.07)	0.87(0.05)	0.86(0.06)	0.88(**0.04**)	0.85(0.06)	0.79(0.10)	0.84(0.06)	0.86(0.07)
NSS	Without CB	**3.38**(0.78)	1.08(0.54)	1.68(**0.47**)	1.00(0.50)	2.36(0.73)	1.17(0.60)	0.71(0.74)	1.64(0.87)	1.38(0.73)
	With CB	**3.41**(0.78)	1.33(0.55)	1.82(0.50)	1.40(**0.32**)	2.42(0.71)	1.56(0.59)	1.09(0.73)	1.60(0.93)	1.74(0.71)
CC	Without CB	**0.80(0.08)**	0.29(0.13)	0.46(0.09)	0.28(0.13)	0.63(0.10)	0.29(0.14)	0.19(0.20)	0.46(0.21)	0.37(0.18)
	With CB	**0.82(0.07)**	0.39(0.13)	0.53(0.10)	0.42(**0.07**)	0.68(0.09)	0.41(0.13)	0.32(0.19)	0.47(0.23)	0.49(0.17)

5. Conclusions

For saliency detection on face images, we have proposed in this paper a learning-based method to take into account the top-down channels of face and facial features. To facilitate saliency analysis of face images, we first established an eye tracking database including 510 face images. Working on our database, GMMs were learnt from the training fixations to model the top-down saliency distributions within face. Combining our GMM-based top-down features (i.e., face and facial feature) with the conventional bottom-up features (i.e., color, intensity, and orientation), our saliency detection method can accurately predict human visual attention on face images. Moreover, weights corresponding to top-down features were optimized by learning the relationship between the weights and face size, since the amount of visual attention on face is relevant to the face size. Finally, experimental results validated that our method significantly advanced saliency detection on face images, as our method drastically outperformed other 8 state-of-the-art methods, in terms of AUC, CC, and NSS.

Acknowledgement. We would like to thank Lai Jiang and Zhaoting Ye for their valuable discussion and help in the evaluations. Also, we would like to thank KingFar International Inc to provide the eye tracker and its technical support. This work was supported by the NSFC projects under Grants 61573037, 61202139, and 61471022, the China 973 program under Grant 2013CB329006, and the MSRA visiting young faculty program.

References

[1] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207, 2013. 1, 6

[2] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, pages 921–928, 2013. 7

[3] N. J. Butko and J. R. Movellan. Optimal scanning for faster object detection. In *CVPR*, pages 2751–2758, 2009. 1

[4] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NIPS*, pages 241–248, 2008. 1, 2, 5, 6, 8

[5] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. In *CVPR*, pages 473–480. IEEE, 2011. 1, 6, 8

[6] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya. Visual attention in quality assessment. *Signal Processing Magazine, IEEE*, 28(6):50–59, 2011. 1

[7] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11, 2013. 1, 6, 8

[8] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):989–1005, 2009. 1

[9] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. 2008. 5

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 1

[11] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):194–201, 2012. 1, 6, 8

[12] Y. Hua, Z. Zhao, H. Tian, X. Guo, and A. Cai. A probabilistic saliency model with memory-guided top-down cues for free-viewing. In *ICME*, pages 1–6, 2013. 1

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998. 1, 4, 6, 8

[14] M. Jiang, J. Xu, and Q. Zhao. Saliency in crowd. In *ECCV*, 2014. 1

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 1, 2, 6, 8

[16] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988. 6

[17] E. Matin. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899, 1974. 1

[18] T. K. Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996. 2, 4

[19] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005. 6

[20] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. In *ACM transactions on graphics (TOG)*, volume 29, page 160. ACM, 2010. 1

[21] J. Saragihand, S. S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, pages 1034–1041, 2009. 3, 6

[22] A. Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003. 1

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *ICCV*, volume 1, pages I–511, 2001. 1

[24] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, 2006. 1, 5

[25] M. Xu, X. Deng, S. Li, and Z. Wang. Region-of-interest based conversational hevc coding with hierarchical perception model of face. *Selected Topics on Signal Processing, IEEE Journal of*, 8(3):475–489, 2014. 1

[26] J. Zhang and S. Sclaroff. Saliency detection: a boolean map approach. In *ICCV*, pages 153–160, 2013. 1, 6, 8

[27] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3):9, 2011. 1, 2, 5, 6, 7, 8

[28] G. Zhu, Q. Wang, and Y. Yuan. Tag-saliency: Combining bottom-up and top-down information for saliency detection. *Computer Vision and Image Understanding*, 118:40–49, 2014. 1