Learning-based Saliency Detection of Face Images

Yun Ren, *Student Member, IEEE,* Zulin Wang, *Member, IEEE,* and Mai Xu, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel method to detect saliency on face images. In our method, face and facial features are extracted as two top-down feature channels, linearly integrated with three traditional bottom-up features of color, intensity, and orientation, to yield final saliency map of a face image. By conducting an eye tracking experiment, a database with human fixations on 510 face images are obtained for analyzing the fixation distribution on face region. We find that fixations on face regions can be well modeled by Gaussian mixture model (GMM), corresponding to face and facial features. Accordingly, we model face saliency by Gaussian mixture model (GMM), learnt from the training data of our database. In addition, we investigate that the weights of face feature channels rely on the face size in images, and the relationship between the weights and face size is therefore estimated by learning from the training data of our eye tracking database. The experimental results validate that our learning-based method is capable of dramatically improving the accuracy of saliency detection on face images over other 10 state-of-the-art methods. Finally, we apply our saliency detection method to compress face images, with an improvement on visual quality or saving on bit-rate over the existing image encoder.**

***Index Terms*—Saliency detection, GMM, face image.**

I. Introduction

A. Background

THE study on the human visual system (HVS) [2] reveals that the distribution of human visual attention on a scene is uneven. In other words, most attention is drawn by a small region (namely salient region), whereas little attention focuses on other region (non-salient region). As a result, human is able to survive in everyday tremendous visual data. Predicting the distribution of visual attention on images or videos is an important way to explore how human perceive the world, and saliency detection is such a way to make computers predict the distribution of visual attention. In fact, saliency detection has been extensively applied in many computer vision and image processing areas, like object detection [3], object segmentation [4], visual quality assessment [5], image retrieval [6] and image/video coding [7].

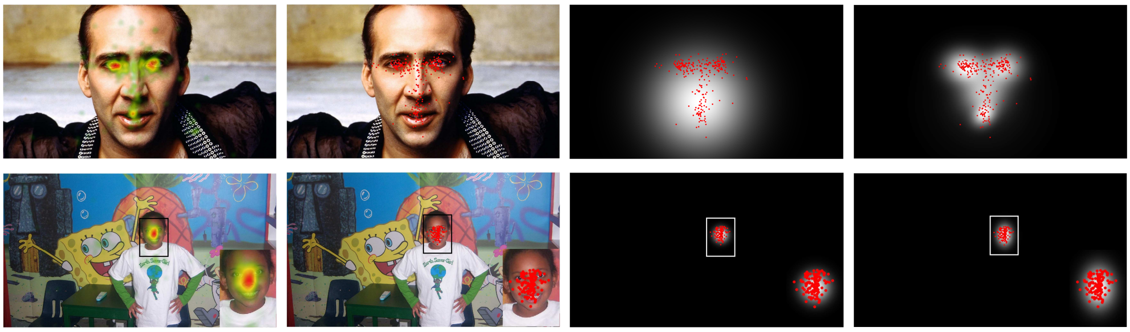
B. Related work

The existing methods for saliency detection can be divided into two classes: either bottom-up or top-down methods. The pioneering bottom-up method on saliency detection is Itti's model [8]. In [8], Itti *et al.* proposed to establish center-surround responses in feature channels of color, intensity and orientation, for yielding the conspicuity maps. Then, the final saliency map is achieved by linearly integrating conspicuity maps of all three feature channels. The fast decade has witnessed extensive advanced work on bottom-up saliency detection, e.g., [9]–[29]. For example, benefitting from 0graph theory, graph-based visual saliency (GBVS) method [12] was proposed to detect the saliency of an image, by forming and then normalizing activation maps, which are also based on the bottom-up features of color, intensity and orientation. Later, the wavelet transform was applied in [13] as a kind of bottom-up features for saliency detection. Taking advantage of the development in signal processing field, some latest signal processing algorithms have been incorporated in saliency detection, e.g., spectral analysis based [15]–[18], sparse representation based [20], [21], and region covariances based [22] methods. Recently, other advanced methods, e.g., online salient dictionary learning (OSDL) [27] and boolean map based saliency (BMS) [28], have also been proposed to predict image saliency. Most recently, some convolutional neural networks (CNN) approaches, such as [23], [24] have been proposed, benefitting from the development of deep learning. In compressed domain, [25] and [29] have been proposed to detect saliency via utilizing bottom-up features of H.264 and H.265 bitstreams, respectively.

In fact, top-down visual features is significant in determining the saliency of images/videos. Thus, the top-down methods of saliency detection have been studied broadly in [30]–[36]. For face images, Cerf *et al.* [32] found that face is a significant top-down cue to draw attention, as their eye tracking experiments shows that faces were fixated on in 88.9% within first two fixations (150 face images viewed by 7 subjects). Therefore, they combined Viola & Jones (VJ) face detector [37] with Itti's model [8], to improve the performance of saliency detection for face images. Since it is more reasonable to learn how important face is on drawing visual attention, some latest methods [38]–[42] have been proposed to employ machine learning algorithms to advance top-down saliency detection of Cerf's work [32]. For instance, Judd *et al.* [38] utilized a classifier to learn several low, middle and high level image features (e.g. intensity, gist features and faces) to combine top-down and bottom-up saliency detection in a single unified formulation. Besides, Zhao [40] quantified the weight of the face channel on drawing attention by utilizing fixations on face images. Lately, Jiang *et al.* [41] extended Cerf's work [32] to detect saliency in a scene composed of multiple faces, i.e., saliency detection in a crowd. Specifically, for detecting saliency in a crowd, Jiang employed multiple kernel learning (MKL) to learn a more robust discrimination between salient and non-salient regions in multi-face scenes. In [42], Ren *et al.* proposed to predict visual attention drawn by different parts of a single face, the sizes of which are fixed in the images, i.e., small, medium and large.

Y. Ren, Z. Wang and M. Xu are with the School of Electronic and Information Engineering, Beihang University, Beijing, 100191 China. M. Xu is the corresponding author of this paper (Email: MaiXu@buaa.edu.cn). This work was supported by NSFC under grant numbers 61202139 and 61573037.

A short version of this paper was presented in ICCV 2015 [1].



(a) Fixation Heatmap (b) Fixations (c) Isotropic GM [40] (d) Learnt GMM

Fig. 1. Learnt GMM vs Isotropic GM for modeling saliency in face region. Both fixations and saliency maps on the face regions are displayed.

C. Our work and main contributions

Although faces have considered for saliency detection in the traditional approaches, these approaches do not explore fixation distribution within faces. As can be seen in Figure 1, the assumption of a simple isotropic Gaussian model (GM) [32], [40] for saliency distribution in face cannot well model visual attention drawn by faces. We can see from this figure that, for images including small faces, the non-isotropic GM is with high accurate in modeling saliency distribution on face. For images including large faces, a single GM is not sufficient, as fixations tend to cluster around the facial features (e.g., eyes). Hence, saliency distribution, in the form of Gaussian mixture model (GMM), needs to be learnt from eye fixations on face images. Figure 1-(d) illustrates that saliency with the learnt GMM distribution is more consistent with the ground truth attention. To be more specific, one non-isotropic Gaussian component needs to be modeled for images with small faces, while more than one component can be modeled for images with large-face. Despite GMM already being incorporated in face saliency detection, it merely concentrates on the GMM of fix-sized faces, such that it cannot be used to detect saliency of images with face at various sizes. Thus, we propose in this paper a learning-based method for saliency detection, and it learns various GMMs and the corresponding weights across different face sizes[1], to predict visual attention on free-viewing face images.

This paper is an extended version of our conference paper [1]. Beyond [1], this paper investigates the improvement of our method for saliency detection of faces at various sizes. The generalization of our method is validated by implementing our method on our database and two other public databases, rather than only testing on our database in [1]. Moreover, this paper provides a potential application of our method in JPEG-based face image compression. The main contributions of this paper are listed as follows:

- We establish a large-scale eye tracking database for visual attention analysis on face images. The ground truth fixations and images of our database are available online[2].
- We model human visual attention attended to face regions using learnt GMM distribution, the weights of which are also learnt with respect to face sizes.
- We apply our saliency detection method to the task of face image compression, which enables bit-rate saving or quality improvement over existing image encoders.

D. Organization of this paper

The rest of this paper is organized as follows. In Section II, we introduce our eye tracking database with a detailed analysis. Then, Section III proposes a novel method for saliency detection on face images. Section IV shows the experimental results to validate our proposed method. Afterwards, Section V discusses an application of our saliency detection method. Finally, Section VI concludes this paper.

II. Database and Analysis

Upon face images, face is an obvious top-down feature for drawing visual attention. Furthermore, intuitively facial features (i.e. eyes, nose, and mouth) may attract more human fixations than other regions in face. To verify this insight, we conducted a database of face images and analyzed the fixation distribution in our database. This section introduces our database and analysis. Specifically, Section II-A discusses the details in database conduction, and Section II-B lists the analysis results on fixations in our database.

A. Database

For the analysis of visual attention on face images, we conducted an eye tracking experiment to establish a database containing extensive fixations on various face images. To be more specific, we follow the below steps to set up our database:

Firstly, we randomly selected 510 face images from Google. There are two criteria when selecting images: (1) The original resolution of images is 1920×1080. (2) All images contain only one frontal face (turning degree of head $< 45°$. As a result, we found that the face sizes in all images vary from 0.0016 to 0.3018. Figure 2 illustrates all face sizes of 510 images. In this figure, we can see that face sizes are distributed in log function.

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

[2] The database is available on our website: $https://github.com/RenYun2016/face$.

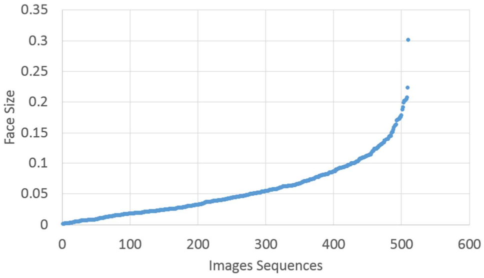


Fig. 2. Face sizes of all 500 images in our database.

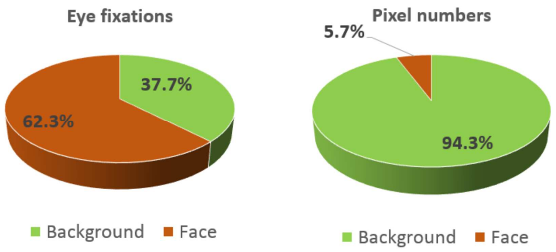


Fig. 3. Eye fixations and pixel numbers belonging to face regions and background regions.

Secondly, 14 male and 10 female, aging from 19 to 35, were employed as volunteers to observe 510 face images in our eye tracking experiment. Note that among 24 subjects there were two subjects having research experience of saliency detection, and other 22 subjects without any background in saliency detection were native to the purpose of the eye tracking experiment.

Thirdly, we conducted our eye tracking experiment by utilizing a Tobii TX300 eye tracker to record the fixations of 24 subjects. Tobii TX300 was applied at a sample rate of 300 Hz, with the same resolution (1920×1080) as face images. The distance between subjects and the monitor of the eye tracker is around 60 cm. Thus, the visual angle of the stimuli was about $26.8^\circ \times 46.0^\circ$. Besides, all subjects were required to have 9-point calibrations for the eye tracker before the experiment. During the experiment, all subjects look at images with free-view mode. Note that all images are displayed randomly, and a 2-second displaying of black image was inserted to correct a drift.

After the experiment, we collected 151,511 fixations for all 510 images, averagely about 300 fixations for each image. Note that we provide all 510 face images, eye tracking data, and corresponding Matlab code on $https://github.com/RenYun2016/face$ for further saliency detection research.

B. Database analysis

Based on the above eye tracking database, we analyze the distribution of visual attention on face and facial features. All 510 face images containing 151,511 fixations in the database are included in the statistical analysis. It should be noted his paper applies a real-time face alignment method [43] to extract face and facial features from face images, obeying the point distribution model (PDM) with 66 landmark points. Through the statistical analysis, we have obtained the following observations:

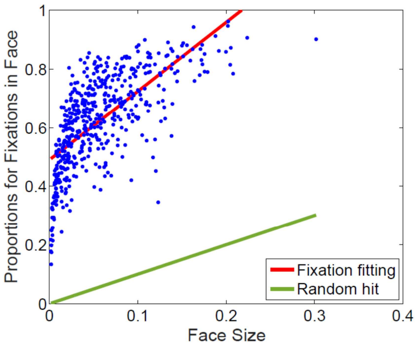


Fig. 4. Proportions for fixations in face. The blue points are proportions for fixations in face of all 510 images. Besides, the red line stands for the linear fitting curve of all blue points. The green line is the probability that points randomly fall into the face region, which is similar to the proportion of face region to the whole image.

Observation 1: Face attracts significantly more visual attention than background.

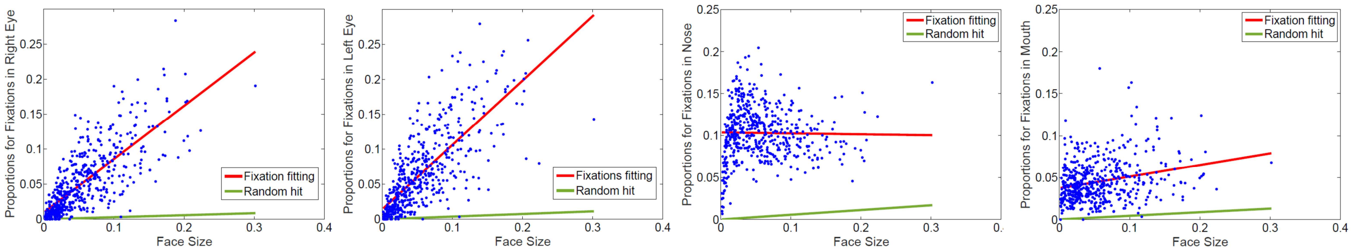
Here, we count fixations of all 510 face images falling into face and background, and list the results in Figure 3. Besides, we also count the pixels belonging to face and background. From Figure 3, we can see that although faces averagely take up only 5.7% of whole images, they draw 62.3% of eye fixations. The results imply that face is more important than background on attracting visual attention. This completes the analysis of Observation 1.

Observation 2: Face attracts more visual attention when face size enlarged.

It is an insight that visual attention on face increases when face size is enlarged. To validate such a insight, we plot in Figure 4 the proportions of fixations on faces versus face sizes, for all 510 images in our database. As shown in Figure 4, the proportions of fixations on faces grow when face size is increased. Besides, all points for proportion of fixations on face are above the curve of random hit. The random hit curve is defined as the probability that a fixation falls into the region of face at random. This again shows that face is with rather large saliency in an image. Besides, one may see from Figure 4 that the increment of fixation fitting curve is much faster than that of the random hit, alongside the enlarged face size. Therefore, we can conclude that much more attention is paid to face once the face is viewed at a large size. This completes the analysis of Observation 2.

Observation 3: Visual attention on eyes and mouth increases along with the enlarged size of face in videos, whereas the attention on nose is invariant to the face size.

Compared with other region in face, facial features have more complex textures. Thus, it is intuitive that visual attention on facial features, i.e., left eye, right eye, nose, and mouth, may be much more than that on other region in face, when the image is displayed with a close up view of face. Then, we investigate the visual attention on facial features by statistical analysis on distribution of the eye fixations within face regions in our database. Figure 5 shows the proportions of fixations on each facial feature versus face sizes, over all 510 images.



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

Fig. 5. Proportions for fixations in facial features. The blue points are proportions for fixations in facial features of all 510 images. Besides, the red line stands for the linear fitting curve of all blue points. The green line is the probability that points randomly falling into the facial feature region, same as the proportion of each facial feature region to the whole image.

We can find out from this figure 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 to large size. However, proportion of fixations on nose region is not grown. It is probably due to the fact that the visual attention shifts from face center (i.e., nose) to other facial features, such as eyes. This completes the analysis of Observation 3.

Based on the three observations, both face and facial features should be taken into consideration in saliency detection, and the weights corresponding to these two channels needs to be correlated with face sizes. Next, we propose our saliency detection method, which mainly focuses on modeling fixation distribution over face and facial features.

III. The Proposed Method

This section presents the proposed method for modeling saliency on face and facial features. In Section III-A, we propose to learn GMM from the training fixations. Next, we present in Section III-B the saliency detection method, which is based on our learnt GMMs. Finally, we propose in Section III-C the algorithm for obtaining optimal weights.

A. Learning GMM

According to our observations, the facial features draw a large amount of visual attention, when 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. First of all, we follow [1] to calibrate and normalize positions of fixations across different face images within a uniformed coordinate system. 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:

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

and

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

where π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are the mixing proportion, mean, and variance of the k-th Gaussian component. In (1), 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 [44] 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 IV.

B. 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 (1) and (2). However, for saliency detection the mean values $\boldsymbol{\mu_k}$ in (1) and (2) are replaced by the central points of 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 [32], 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

$$\mathbf{M} = w_C\mathcal{C}(\mathbf{C}) + w_I\mathcal{C}(\mathbf{I}) + w_O\mathcal{C}(\mathbf{O}) + w_F\mathcal{C}(\mathbf{F}) + w_G\mathcal{C}(\mathbf{G}), \tag{3}$$

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 [11], 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 6 shows an example of overall procedure on our learning-based saliency detection method.

C. Learning optimal weights

Now, the remaining task for saliency detection with (3) 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

TABLE I
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	$\begin{pmatrix} 0.007 & 0.001 \\ 0.001 & 0.009 \end{pmatrix}$	$\begin{pmatrix} 0.013 & -0.002 \\ -0.002 & 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.035 & 0.003 \\ 0.003 & 0.032 \end{pmatrix}$	$\begin{pmatrix} 0.011 & -0.001 \\ -0.001 & 0.033 \end{pmatrix}$

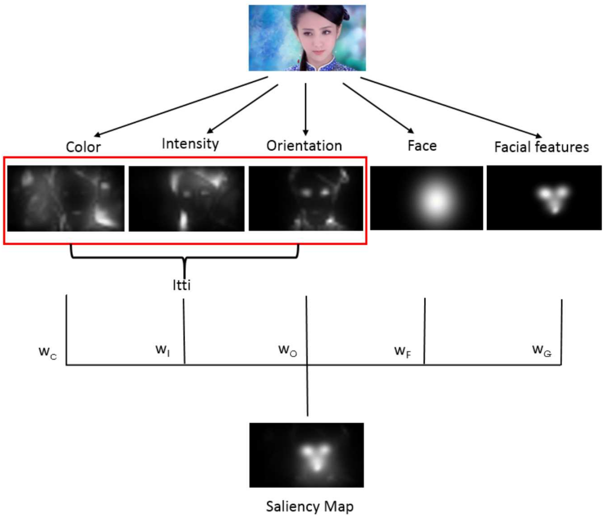


Fig. 6. Framework of our learning-based saliency detection method.

way of [40] to obtain weights $\mathbf{w}$ for each training image, by solving the following ℓ_2-norm optimization formulation:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{V}\mathbf{w} - \mathbf{m}_h\|_2, \quad s.t. \quad ||\mathbf{w}||_1 = 1, \mathbf{w} \geq 0, \tag{4}$$

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, (4) is solved to obtain an optimal weight $\mathbf{w}$ corresponding to this image. To solve (4), the disciplined convex programming approach [45] 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 [40] 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 training images. Specifically, we find that the optimal weights are dependent on face sizes. This is also consistent with the observation of Section II-B, 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,

$$w_F(s) = \sum_{i=0}^{I} a_i s^i, \tag{5}$$

and

$$w_G(s) = \sum_{i=0}^{I} b_i s^i, \tag{6}$$

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 IV, $I = 4$ 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 IV.

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 (5) and (6), 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 IV. Finally, the saliency map of a face image can be worked out via (3) with the learnt optimal weights.

IV. EXPERIMENTAL RESULTS

This section discusses experimental results for both training and test in our method. Specifically, Section IV-A shows the learnt results on the learnt GMMs and weights in training set. Sections IV-B and IV-C provide testing results of our method and other 10 state-of-the-art methods evaluated on our and other databases. In our experiments, the saliency evaluation metrics of the area under ROC curve (AUC), normalized scanpath saliency (NSS) [46], and linear correlation coefficient (CC) [47] are utilized on all test images. Besides, the saliency maps of several test images are also provided for the comparison. Finally, Section IV-D analyzes the improvement of our method for images with different face sizes.

A. *Training results*

In the experiment, all 510 face images in our database are divided into two groups: training and test sets, without any overlap between the two groups. To be more specific, the training set contains 360 images with 106,067 fixations and the other 150 images with 45,444 fixations are included in test set.

Learnt GMMs. Based on the training set of 360 images, we obtain the learnt GMMs for top-down feature channels of saliency detection by utilizing the method in Section III-A. For the face channel, we learn the GMM with only one Gaussian component. Note that 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 [43]), 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

$$\Sigma_1 = \begin{pmatrix} 0.024 & 0 \\ 0 & 0.039 \end{pmatrix}.$$

As can be seen above, there exists the anisotropy in learnt GMMs, rather than the assumption on isotropy of Gaussian distribution in [32].

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 7 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 II-B 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 I.

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 III-C. As aforementioned, the optimal weights w_F and w_G for face and facial feature channels depend on the face size. Figures 8-(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 8-(a) and -(b). We further show in Figure 8-(c) the Pearson's correlation coefficient (PCC) [48] 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 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 (5) 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 (6) 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 (3), with the aforementioned GMMs and optimal weights.

B. Testing results on our database

AUC, CC and NSS. In order to evaluate the performance of saliency detection, we tabulate in Table II the AUC, CC and NSS results of our and other 10 methods. Note that methods with a larger AUC and NSS value, a CC value closer to +1/-1, can better predict the human fixations. In Table II, the evaluation 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 [14]. As seen from Table II, the methods with top-down features, i.e., Cerf *et al.* [32], Judd *et al.* [38], Zhao *et al.* [40], Jiang [41] 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. Note that the latest Jiang [41] performs worst among all top-down methods, due to the fact that it deals with multiple faces rather than single face. Furthermore, our method outperforms other 10 state-of-the-art top-down and bottom-up methods in terms of AUC, CC and NSS. Specially, without CB modeling, there is 0.02 improvement of AUC, 1.02 increase of NSS and 0.17 enhancement of CC over Zhao *et al.* [40], which also integrates the top-down face channel and learns its corresponding weight from training data. The possible reasons for our method outperforming Zhao *et al.* [40] are 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. Obviously, our method is superior to other methods.

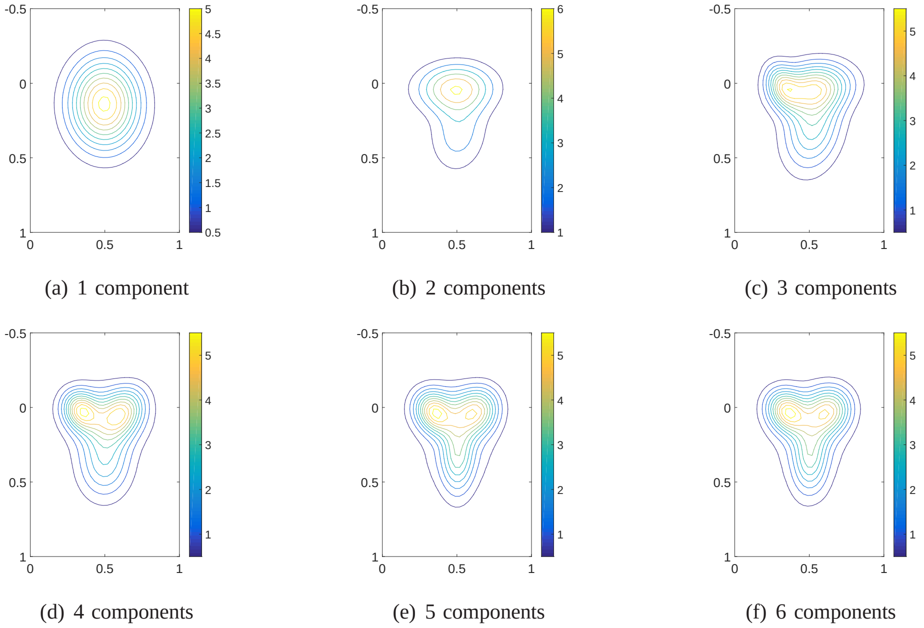
Saliency map. We show the saliency maps of our and other 10 methods in Figure 9. As this figure shows, the saliency map of our method is much more closer to the distribution of human fixations than other 10 methods. Specifically, in our method, when the face size is small (i.e., from first to fourth rows), the face channel in the form of learnt non-isotropic Gaussian works more than other channels, as it is closer to the distribution of human fixations. Similarly, when the face size is large (i.e., from fifth to tenth rows), the success of our saliency results mostly attributes to the channel of facial features, which modeled by the learnt GMMs. 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.

C. Testing results on other databases

In order to test the generalization of our method, we compared our and other 10 methods on detecting saliency of all qualified face images[3] from MIT and NUSEF databases. Note that the model parameters of Section IV-A were utilized here, and they were trained on our database. Similar to Section IV-B, we use AUC, NSS, and CC metrics to evaluate the saliency detection results of all methods. They are reported in Tables III and IV for the scenarios with and without CB, respectively.

As seen from Table III, with the same CB, our method enjoys at least 0.01, 0.19, and 0.04 improvement in AUC, NSS, and CC, for MIT database. For NUSEF database, there exists at least 0.01, 0.23, and 0.06 enhancement in AUC, NSS, and CC. Similar enhancement can be found from Table IV for the case without any CB. Despite being trained on our database, our method still has good performance on other two databases. This verifies the effectiveness of our method in the generalization.

[3] In other databases, all images that contain one frontal face at high quality were selected as qualified images for the test. In this case, 55 and 107 face images were chosen from in MIT and NUSEF databases, respectively, and they are available online: $https://github.com/RenYun2016/face$.



(a) 1 component

(b) 2 components

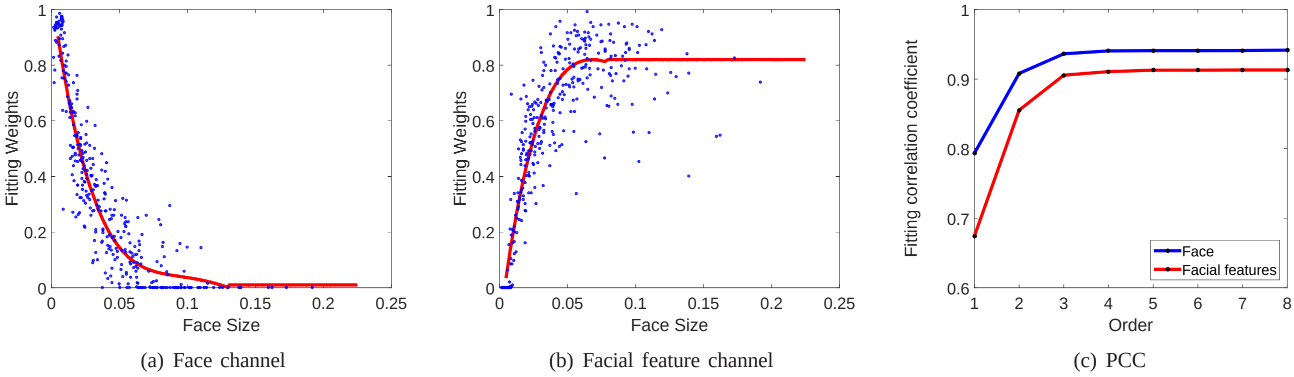
(c) 3 components

(d) 4 components

(e) 5 components

(f) 6 components

Fig. 7. Contours of learnt GMMs with various numbers of Gaussian components.



(a) Face channel

(b) Facial feature channel

(c) PCC

Fig. 8. Fitting weights for face and facial feature channels. Blue dots are the optimal weights of face or facial feature channel in all training images. The red lines in (a) and (b) are the fitting curves of blue dots. (c) shows the PCC values in various orders of polynomial fitting curves.

TABLE II
THE MEAN VALUES (STANDARD DEVIATION) COMPARISON OF OUR AND OTHER METHODS OVER OUR DATABASE

Metric	CB Model	Our	Itti[8]	Cerf[32]	Judd[38]	Zhao[40]	Duan[14]	Hou[18]	Erdem[22]	Zhang[28]	Jiang[41]	OBDL[25]
AUC	*Non-CB*	**0.90(0.04)**	0.52(0.07)	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)	0.78(0.11)	0.75(0.15)
	With CB	**0.90(0.04)**	0.77(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)	0.83(0.07)	0.81(0.10)
NSS	*Non-CB*	**3.38**(0.78)	0.12(0.41)	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)	1.25(0.72)	1.01(0.95)
	With CB	**3.41**(0.78)	0.31(0.51)	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)	1.52(0.70)	0.82(0.81)
CC	*Non-CB*	**0.80(0.08)**	0.02(0.09)	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)	0.36(0.19)	0.26(0.24)
	With CB	**0.82(0.07)**	0.08(0.12)	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)	0.45(0.18)	0.37(0.22)

TABLE III
THE MEAN VALUES (STANDARD DEVIATION) COMPARISON OF OUR AND OTHER METHODS WITH CB MODELING, OVER MIT AND NUSEF DATABASES

Database	Metric	Our	Itti[8]	Cerf[32]	Judd[38]	Zhao[40]	Duan[14]	Hou[18]	Erdem[22]	Zhang[28]	Jiang[41]	OBDL[25]
MIT	*AUC*	**0.85**(**0.05**)	0.74(0.07)	0.83(0.06)	0.81(0.07)	0.84(**0.05**)	0.80(0.07)	0.76(0.09)	0.81(0.07)	0.84(0.06)	0.80(0.06)	0.79(0.09)
	NSS	**2.57**(0.72)	0.47(0.70)	1.65(0.55)	1.19(**0.33**)	2.38(0.70)	1.39(0.57)	1.03(0.70)	1.42(0.75)	1.61(0.62)	1.17(0.45)	1.25(0.63)
	CC	**0.72**(0.12)	0.14(0.17)	0.52(0.14)	0.41(**0.10**)	0.68(0.12)	0.46(0.16)	0.34(0.20)	0.46(0.20)	0.51(0.16)	0.45(0.22)	0.43(0.23)
NUSEF	*AUC*	**0.83**(**0.05**)	0.72(0.06)	0.81(**0.05**)	0.80(0.06)	0.82(**0.05**)	0.80(0.07)	0.75(0.08)	0.80(0.06)	0.81(0.06)	0.79(0.07)	0.79(0.07)
	NSS	**1.94**(0.56)	0.21(0.38)	1.42(0.38)	1.17(**0.31**)	1.71(0.50)	1.33(0.52)	0.95(0.54)	1.25(0.65)	1.39(0.48)	1.31(0.57)	1.27(0.56)
	CC	**0.75**(0.11)	0.08(0.11)	0.60(0.12)	0.51(**0.10**)	0.69(0.12)	0.56(0.16)	0.40(0.20)	0.52(0.21)	0.56(0.13)	0.46(0.22)	0.41(0.21)
Average	*AUC*	**0.84**(**0.05**)	0.73(0.06)	0.82(0.06)	0.80(0.06)	0.83(**0.05**)	0.80(0.07)	0.76(0.09)	0.81(0.06)	0.82(0.06)	0.79(0.07)	0.79(0.08)
	NSS	**2.14**(0.68)	0.29(0.48)	1.49(0.45)	1.18(**0.32**)	1.93(0.65)	1.35(0.54)	0.97(0.59)	1.31(0.69)	1.46(0.54)	1.26(0.53)	1.27(0.58)
	CC	**0.74**(**0.11**)	0.10(0.13)	0.58(0.13)	0.48(**0.11**)	0.69(0.12)	0.53(0.17)	0.38(0.20)	0.50(0.21)	0.55(0.14)	0.45(0.22)	0.42(0.21)

TABLE IV
THE MEAN VALUES (STANDARD DEVIATION) COMPARISON OF OUR AND OTHER METHODS WITHOUT CB MODELING, OVER MIT AND NUSEF DATABASES

Database	Metric	Our	Itti[8]	Cerf[32]	Judd[38]	Zhao[40]	Duan[14]	Hou[18]	Erdem[22]	Zhang[28]	Jiang[41]	OBDL[25]
MIT	*AUC*	**0.85**(0.06)	0.64(0.09)	0.82(0.07)	0.76(0.09)	0.84(**0.05**)	0.81(0.06)	0.50(0.21)	0.79(0.08)	0.81(0.07)	0.76(0.09)	0.74(0.13)
	NSS	**2.57**(0.70)	0.56(0.73)	1.58(0.50)	0.99(**0.44**)	2.39(0.69)	1.36(0.49)	0.74(0.78)	0.92(0.59)	1.35(0.61)	1.02(0.50)	1.00(0.71)
	CC	**0.72**(**0.12**)	0.13(0.17)	0.50(0.13)	0.32(0.14)	0.68(0.13)	0.44(0.13)	0.24(0.23)	0.30(0.18)	0.42(0.17)	0.33(0.22)	0.27(0.23)
NUSEF	*AUC*	**0.82**(**0.05**)	0.64(0.10)	0.80(**0.05**)	0.73(0.08)	0.81(**0.05**)	0.8(0.06)	0.44(0.21)	0.78(0.07)	0.77(0.09)	0.76(0.08)	0.75(0.10)
	NSS	**1.91**(0.56)	0.28(0.43)	1.33(**0.35**)	0.92(0.39)	1.67(0.52)	1.30(0.41)	0.65(0.54)	0.85(0.48)	1.15(0.52)	1.17(0.56)	1.01(0.63)
	CC	**0.73**(**0.12**)	0.07(0.10)	0.56(0.13)	0.38(0.15)	0.66(0.13)	0.55(0.13)	0.27(0.22)	0.35(0.19)	0.46(0.18)	0.35(0.22)	0.28(0.22)
Average	*AUC*	**0.83**(**0.06**)	0.64(0.10)	0.81(**0.06**)	0.74(0.09)	0.82(**0.06**)	0.80(**0.06**)	0.46(0.21)	0.78(0.07)	0.78(0.09)	0.76(0.09)	0.74(0.11)
	NSS	**2.12**(0.68)	0.37(0.53)	1.41(0.42)	0.94(**0.41**)	1.90(0.67)	1.32(0.43)	0.68(0.63)	0.87(0.52)	1.21(0.56)	1.12(0.54)	1.00(0.65)
	CC	**0.72**(**0.12**)	0.09(0.13)	0.54(0.13)	0.36(0.15)	0.67(0.13)	0.52(0.14)	0.26(0.22)	0.34(0.19)	0.45(0.17)	0.34(0.22)	0.28(0.23)

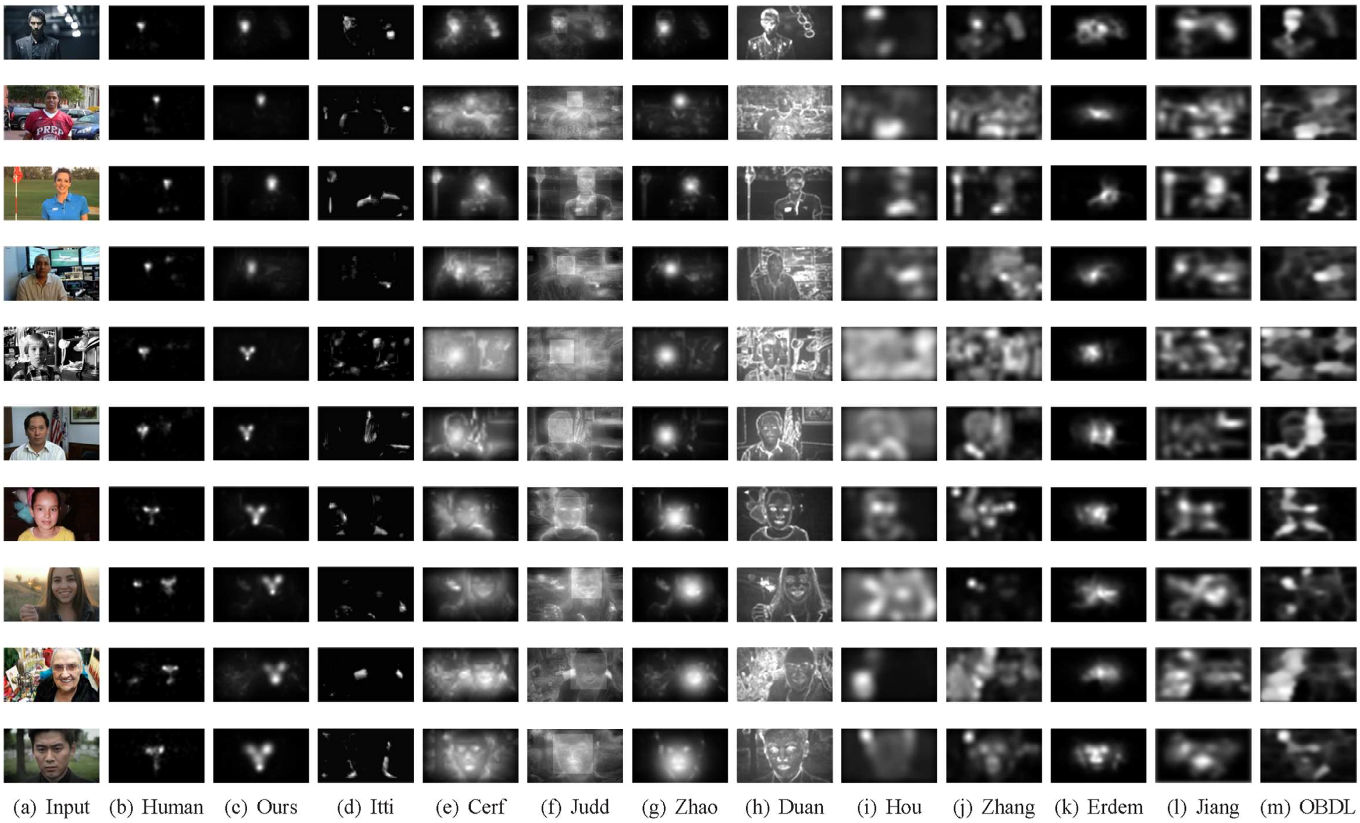

(a) Input (b) Human (c) Ours (d) Itti (e) Cerf (f) Judd (g) Zhao (h) Duan (i) Hou (j) Zhang (k) Erdem (l) Jiang (m) OBDL

Fig. 9. Saliency maps of our method and other state-of-the-art methods without any CB.

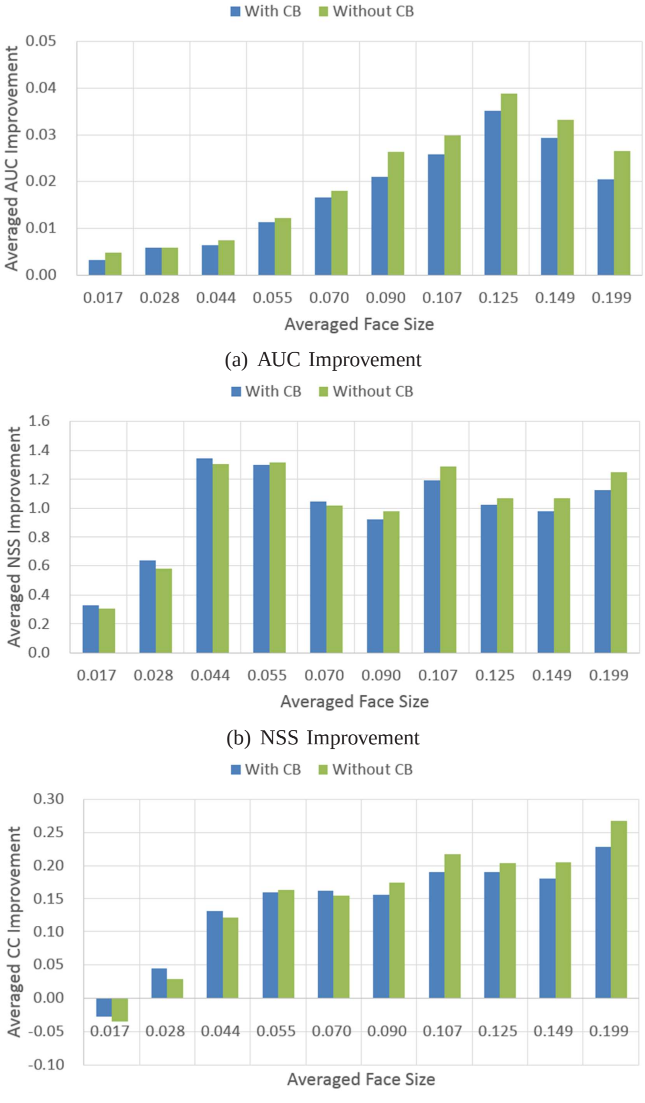
D. Improvement analysis of our method

In this subsection, we focus on investigating the saliency detection improvement of our method at different face sizes. Since Section IV-B has shown that Zhao *et al.* [40] performs the best except our method, we use [40] as an anchor for the improvement analysis of our method. Here, all 150 test images of Section IV-B were used for our analysis. These 150 images were first sorted by increased face sizes, and then they were averagely divided into 10 groups according to the sizes of their corresponding faces. As a result, each group contains 15 images with similar-sized faces. Finally, the averaged face sizes and improvement of each group (in terms of AUC, NSS, and CC) were calculated for the analysis.

Figure 10 shows the averaged improvement of saliency detection accuracy over Zhao *et al.* [40] for the 10 groups of images, the averaged face sizes of which monotonically increase. From this figure, we can see that the greater improvement can be achieved in our method, when the face size is large (generally over 0.044) in the image. This implies that our method performs better when images are with large-sized faces. Such better performance is probably due to the effectiveness of GMM modeling on facial feature channel, as Figure 8 shows that the weight of facial feature channel roughly arrives at maximal value once the face size is larger than 0.05.

V. IMAGE COMPRESSION APPLICATION

In fact, saliency detection has potential to be applied in some computer vision and image processing tasks, such as image compression, image retargeting and visual quality assessment. In this section, we presented a simple application of our saliency detection method to face image compression. Some advanced approaches may be also developed on the basis of our saliency detection method, for the further improvement.



(a) AUC Improvement

(b) NSS Improvement

(c) CC Improvement

Fig. 10. AUC, NSS, and CC improvement of our method over Zhao *et al.* [40], along with increased face sizes. Note that the averaged results are reported here for 10 groups of images, each of which contains faces at similar sizes. Specifically, the horizontal axes stand for the face size averaged in each group. The vertical axes mean the averaged accuracy improvement of our method over [40], in (a) AUC, (b) NSS, and (c) CC. Note that the results of saliency detection with CB and without CB are both shown in this figure.

However, it is out of the scope of this paper, as we only focus on the potential of our saliency detection method in the application of face image compression.

A. Approach for face image compression

Figure 11 summarizes the framework of our approach for face image compression. As can be seen from this figure, a preprocess step is integrated into the conventional Joint Photographic Experts Group (JPEG) approach. In the preprocess step, our saliency detection method is applied to generate the saliency map of the input face image. Then, the input image is preprocessed by an LPF, the cutoff frequency of which is determined by the saliency map. That is the non-salient regions correspond to low cutoff frequency in the LPF, such that the detailed texture is removed in these regions. Finally, the preprocessed image is compressed by the JPEG encoder. As a result, the texture details are reduced in non-salient regions with little overall quality degradation, as these regions can hardly attract attention. In return, the bit-rates of compressed images can be saved or the quality of salient blocks (e.g., face and facial features) can be enhanced using saved bit-rates. Since the preprocess of our approach is independent of the encoder, it can be easily transplanted into other state-of-the-art encoder[4].

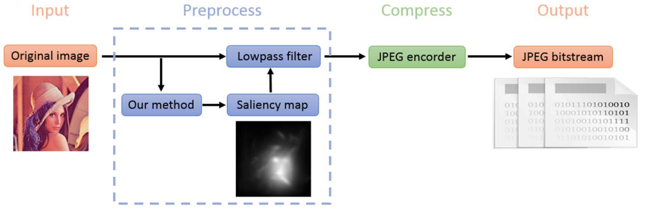


Fig. 11. The framework of our approach for face image compression, which is based on our saliency detection method and the conventional JPEG encoder.

The same as JPEG, the input image (denoted by $\mathbf{P}$) is divided into N non-overlapping blocks with 8×8 pixels, i.e., $\mathbf{P} = \{\mathbf{p}_n\}_{n=1}^{N}$. Mathematically, each block $\mathbf{p}_n$ can be processed to be $\mathbf{p}_n^*$ by

$$\mathbf{p}_n^* = \mathcal{IDCT}(\mathrm{LPF}(\mathcal{DCT}(\mathbf{p}_n))), \tag{7}$$

where $\mathcal{DCT}$ and $\mathcal{IDCT}$ denote the DCT and inverse DCT, respectively. In (7), LPF$(\cdot)$ is an LPF as follows,

$$d_n^*(i,j) = \begin{cases} d_n(i,j) & i+j \leq T_n \\ 0 & i+j > T_n, \end{cases} \quad s.t. \quad 1 \leq i,j \leq 16, \tag{8}$$

In (8), $d_n(i,j)$ and $d_n^*(i,j)$ are the (i,j)-th DCT coefficients of the n-th block before and after the LPF. T_n is the cutoff frequency of the LBF for the n-th block, which is proportional to its averaged saliency S_n. For more details about the calculation of T_n, see Appendix A. Finally, the preprocessed image $\mathbf{P}^* = \{\mathbf{p}_n^*\}_{n=1}^{N}$ can be obtained with (7) and (8), and then compressed by the conventional JPEG encoder.

B. Results of face image compression

Since our approach simply adds a preprocess step before the JPEG encoder, we compare the results of our and conventional JPEG approaches on compressing ten face images. These ten images include all color face images from [49] (i.e., Lena, Tiffny, Girl01, Girl02 and Girl03), and five face images randomly selected from our test set of Section IV (as shown in Figure 12). Here, DMOS is used as the metric for evaluating the subjective visual quality of compressed images. See Appendix B for our subjective experiment on obtaining the DMOS value of each compressed image.

First, we report in Table V the visual quality improvement of our approach, in terms of the DMOS difference. In general,

[4] Here, the JPEG encoder is used in our application, due to its extremely high popularity. In fact, most of other state-of-the-art image encoders, such as H.265 intra coding, are also block-based approaches. Thus, they can be in conjunction with our saliency detection method for the further improvement of coding efficiency.

TABLE V

DMOS COMPARISON OF OUR AND CONVENTIONAL JPEG APPROACHES FOR COMPRESSING IMAGES AT SAME LOW BIT-RATES

Image	Resolution (pixel)	Bit-rate (bpp)		DMOS		
		JPEG	Our	JPEG	Our	Diff.
Lena	512×512	0.37	0.37	72.0	58.1	-13.9
Tifffny	512×512	0.27	0.27	62.9	57.0	-5.9
Girl01	256×256	0.37	0.37	60.6	53.8	-6.8
Girl02	256×256	0.29	0.29	73.6	66.7	-6.9
Girl03	256×256	0.45	0.45	61.1	56.7	-4.4
Fig.12(a)	1920×1080	0.20	0.20	59.9	52.4	-7.5
Fig.12(b)	1920×1080	0.27	0.27	77.7	58.9	-18.8
Fig.12(b)	1920×1080	0.19	0.19	56.5	50.2	-6.3
Fig.12(d)	1920×1080	0.29	0.29	73.3	54.9	-18.4
Fig.12(e)	1920×1080	0.24	0.23	73.8	65.3	-8.5
Average	-	-	-	-	-	-9.7

TABLE VI

BIT-RATE COMPARISON OF OUR AND CONVENTIONAL JPEG APPROACHES FOR COMPRESSING IMAGES AT THE SIMILAR DMOS

Image	Resolution (pixel)	Bit-rate (bpp)			DMOS		
		JPEG	Our	Saving	JPEG	Our	Diff.
Lena	512×512	7.01	5.32	24.1%	33.9	38.2	$+4.3$
Tifffny	512×512	6.96	6.09	12.5%	40.7	35.9	-4.8
Girl01	256×256	6.98	6.05	13.3%	34.7	35.6	$+0.9$
Girl02	256×256	5.20	4.13	20.1%	39.7	38.6	-1.1
Girl03	256×256	6.56	5.81	11.4%	35.3	40.0	$+4.7$
Fig.12(a)	1920×1080	4.94	3.67	25.7%	33.3	33.4	$+0.1$
Fig.12(b)	1920×1080	5.03	3.09	38.6%	37.7	41.7	$+4.0$
Fig.12(c)	1920×1080	3.88	3.41	12.1%	36.0	33.4	-2.6
Fig.12(d)	1920×1080	4.91	4.12	16.1%	41.8	39.8	-2.0
Fig.12(e)	1920×1080	4.43	3.74	15.6%	39.5	39.5	$+0.0$
Average	-	-	-	19.0%	-	-	$+0.4$

the smaller DMOS means better subjective quality of the compressed image. Thus, Table V shows that at the same (low) bit-rate, the subjective quality of images compressed by our approach is much superior to those compressed by JPEG. Next, Table VI tabulates the bit-rate saving of our approach with JPEG as an anchor. We can see from this table that our approach has 19% bit-rate saving in average, with roughly similar subjective visual quality.

Figure 13 further visualizes two selected images compressed by our and the conventional JPEG approaches, respectively. From these pictures, one may observe that there exist evident block effects and severe visual distortion in face for the images compressed by JPEG. By contrast, our approach has better visual quality in face regions at the expense of quality degradation in other non-salient regions, thus favoring the subjective feeling on compressed images. In summary, our approach is able to improve the performance of the widely used JPEG encoder, which demonstrates an effective application of our saliency detection method.

VI. Conclusions

In this paper, we have proposed a learning-based method for detecting saliency on face images, which utilizes face and facial features as two top-down feature channels to predict saliency. To analyze fixation distribution on face regions, we conducted an experiment to obtain an eye-tracking database composed of 510 face images. Based on the statistical analysis over our database, we proposed to learn GMMs, which are modeled to compute saliency for both face and facial features. Note that the parameters of GMMs were learnt from the eye tracking data of the training set. In our method, the saliency map was computed by the linear combination of GMM-based features (face and facial feature) and the traditional low-level features (color, intensity, and orientation). Besides, we validated that there is a kind of relevance between the weights of the linear combination for each channel and face size in an image, which can also been learnt from the training set. The evaluation results of AUC, CC, and NSS showed that, compared to other 10 state-of-the-art methods, our method predicted saliency with much higher accuracy. Finally, we demonstrated a simple yet effective application of our saliency detection method on face image compression.

There may exist two research directions for the future work. (1) Our work in this paper only considers the frontal faces. In fact, the head pose may also influence the distribution of visual attention on face. Thus, the saliency detection work related to head pose remains to be done, for images with non-frontal faces. (2) Our work in the current form only deals with still face images. In practice, faces are more likely to appear in videos, such as video conferencing scenarios. Thus, extension of our method to conversational videos, incorporating motion information of facial features, shows a promising research trend in future.

Appendix

A. Calculation of cutoff frequency T_n

As aforementioned, the cutoff frequency T_n of the n-th block should be proportional to its averaged saliency S_n. So, T_n may be defined by

$$T_n = \beta \cdot e^{\alpha \cdot S_n}, \tag{9}$$

where $\alpha > 0$ and $\beta > 0$ are two parameters to control the cutoff frequency T_n upon averaged saliency S_n. They can be estimated with the following assumption. For the non-salient blocks with $S_n = 0.001$, we assume that $T_n = 1$. For the salient blocks with $S_n = 0.1$, we assume that $T_n = 15$. Then, α and β can be achieved via solving the following equations:

$$\begin{cases} \beta e^{0.001 \cdot \alpha} = 1 \\ \beta e^{0.1 \cdot \alpha} = 15. \end{cases} \tag{10}$$

After solving the above equations, we have $\alpha = 27.354$ and $\beta = 0.973$. Finally, the cutoff frequency is $T_n = 0.973 \cdot e^{27.354 \cdot S_n}$, for our image compression approach.

B. Subjective assessment for DMOS

DMOS measures subjective quality difference between compressed and uncompressed images. In this paper, it was obtained in the IVQUEST software, by designing the following subjective assessment.

In our subjective assessment, we utilized a single stimulus continuous quality scale (SSCQS) procedure to rate the subjective quality, which is proposed in Rec. ITU-R BT.500 [50]. The scales available for the quality rate are: excellent (100-81), good (80-61), fair (60-41), poor (40-21), and bad (20-1). In our assessment, 10 observers, aging from 19 to 34, were involved to rate the subjective quality of all uncompressed

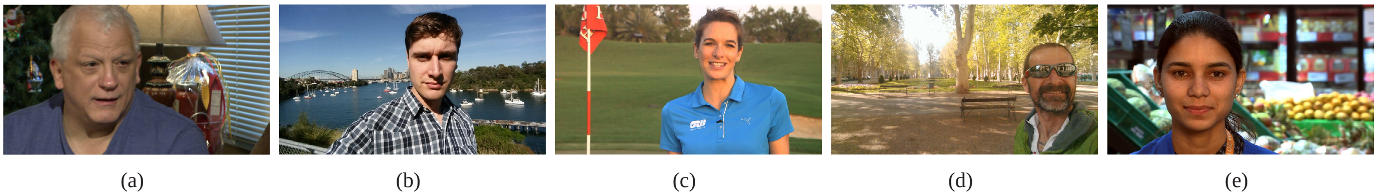
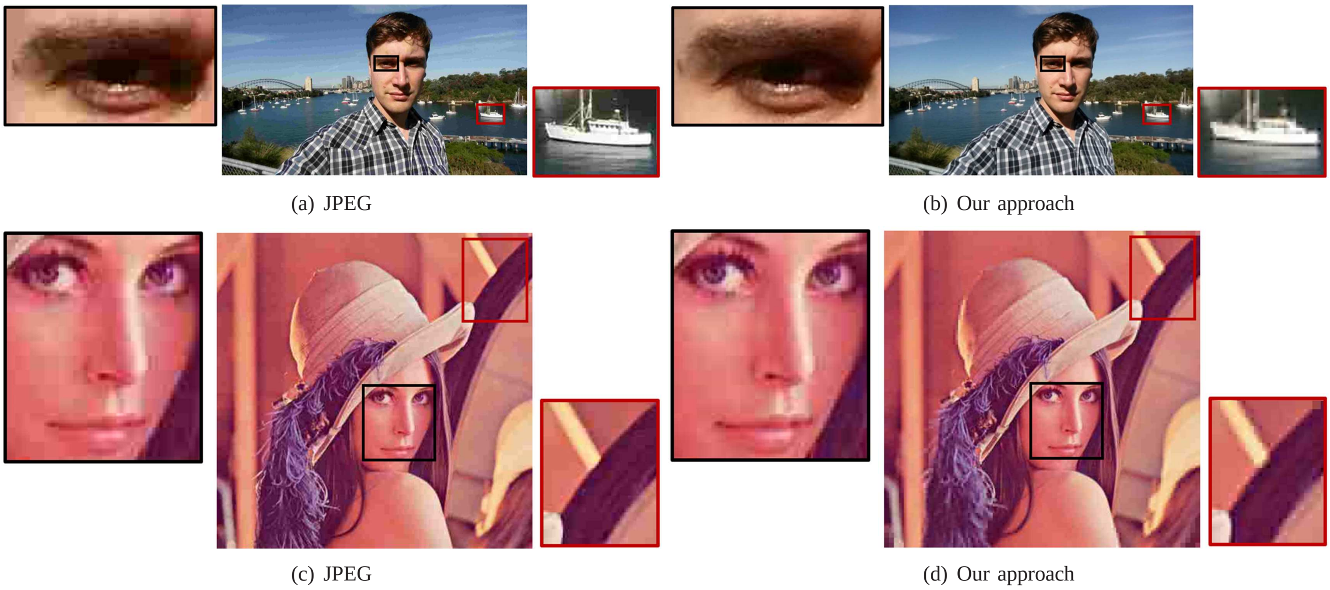

(a) (b) (c) (d) (e)

Fig. 12. Original images used for image compression. Note that these five images were randomly selected from our 150 test images of Section IV.


(a) JPEG

(b) Our approach

(c) JPEG

(d) Our approach

Fig. 13. Visual quality comparison of two compressed images. (a) and (b) show the image of Figure 12(b) compressed by the conventional JPEG approach and our approach at 0.27 bpp. (c) and (d) show the image of Lena compressed by the conventional JPEG approach and our approach at 0.37 bpp.

and compressed images. For the subjective assessment, a 24" SAMSUNG S24B370 LCD monitor was used to randomly display images at their original resolutions. As such, the distortion of scaling operations can be avoided.

Besides, for rational evaluation, the viewing distance for observers was set to be approximately three times of the image height.

After the subjective assessment of all observers, we obtained raw subjective scores of each image. Finally, DMOS for each compressed image is obtained upon those raw scores were converted, using the method in [51]. The DMOS values indicate the subjective visual difference between the uncompressed reference and compressed test images.

References

[1] M. Xu, Y. Ren, and Z. Wang, "Learning to predict saliency on face images," in *ICCV*, 2015.

[2] E. Matin, "Saccadic suppression: a review and an analysis." *Psychological bulletin*, vol. 81, no. 12, p. 899, 1974.

[3] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *CVPR*, 2009, pp. 2751–2758.

[4] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *Multimedia, IEEE Transactions on*, vol. 14, no. 5, pp. 1429–1441, 2012.

[5] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 50–59, 2011.

[6] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," *Multimedia, IEEE Transactions on*, vol. 17, no. 3, pp. 359–369, 2015.

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

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] H. Kim and S. Lee, "Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort," *Multimedia, IEEE Transactions on*, vol. 17, no. 12, pp. 2198–2209, 2015.

[10] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005, pp. 155–162.

[11] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[13] N. İmamoğlu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 96–105, 2013.

[14] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *CVPR*. IEEE, 2011, pp. 473–480.

[15] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, 2010.

[16] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 187–198, 2012.

[17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*. IEEE, 2007, pp. 1–8.

[18] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.

[19] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *Journal of vision*, vol. 12, no. 6, p. 17, 2012.

[20] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, "Regularized feature reconstruction for spatio-temporal saliency detection," *Image Processing, IEEE Transactions on*, vol. 22, no. 8, pp. 3120–3132, 2013.

[21] Y. Luo, Y. Wong, and Q. Zhao, "Label consistent quadratic surrogate model for visual saliency prediction," in *CVPR*, 2015.

[22] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of vision*, vol. 13, no. 4, p. 11, 2013.

[23] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *ICLR*, 2015.

[24] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *ICCV*, 2015, pp. 262–270.

[25] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *CVPR*, 2015, pp. 5501–5510.

[26] J. Lu, X. Wen, H. Shao, Z. Lu, and Y. Chen, "An effective visual saliency detection method based on maximum entropy random walk," in *ICME Workshop*. IEEE, 2016, pp. 1–6.

[27] M. Xu, L. Jiang, Z. Ye, and Z. Wang, "Bottom-up saliency detection with sparse representation of learnt texture atoms," *Pattern Recognition*, vol. 60, pp. 348–360, 2016.

[28] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 889–902, 2016.

[29] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with hevc features," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 369–385, 2017.

[30] A. Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[31] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2049–2056.

[32] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *NIPS*, 2008, pp. 241–248.

[33] S. Chikkerur, T. Serre, and C. Tan, "what and where: A bayesian inference theory of visual attention,," *Vision Research*, vol. 229, no. 4715, p. 782, 2010.

[34] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 5, pp. 523–538, May 2014.

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

[36] B. Ni, M. Xu, T. V. Nguyen, M. Wang, C. Lang, Z. Huang, and S. Yan, "Touch saliency: Characteristics and prediction," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1779–1791, 2014.

[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *ICCV*, vol. 1, 2001, pp. I–511.

[38] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009, pp. 2106–2113.

[39] 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*, 2013, pp. 1–6.

[40] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, p. 9, 2011.

[41] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *ECCV*, 2014.

[42] Y. Ren, M. Xu, R. Pan, and Z. Wang, "Learning gaussian mixture model for saliency detection on face images," in *ICME*. IEEE, 2015, pp. 1–6.

[43] J. Saragihand, S. S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *ICCV*, 2009, pp. 1034–1041.

[44] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[45] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, 2008, pp. 95–110.

[46] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[47] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[48] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[49] A. Weber, http://sipi.usc.edu/database/.

[50] ITU, "Methodology for the subjective assessment of the quality of television pictures," *BT. 500-11, International Telecommunication Union, Geneva, Switzerland*, pp. 53–56, 2002.

[51] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE transactions on*, vol. 19, no. 6, pp. 1427–1441, 2010.