Saliency detection in face videos: A data-driven approach

Mai Xu *IEEE Senior Member*, Yun Ren *IEEE Student Member*, Zulin Wang *IEEE Member*, Jingxian Liu and Xiaoming Tao

***Abstract*—Recently, video conferencing has been popular in multimedia systems, such as FaceTime and Skype. In video conferencing, almost every frame contains a human face. Therefore, it is important to predict human visual attention on face videos by saliency detection, as saliency may be used as a guide to the region-of-interest (ROI) for the content-based applications of face videos. In this paper, we propose a data-driven approach for saliency detection in face videos. From the data-driven perspective, we first establish an eye tracking database that contains fixations of 76 face videos viewed by 40 subjects. Upon the analysis of our database, we find that visual attention is significantly attracted by faces in videos. More importantly, the attention distribution within face regions varies with regard to mouth movement. Since previous works have investigated that it is efficient to model face saliency in still images using a Gaussian mixture model (GMM), the variation of visual attention in videos can be modeled by dynamic GMM (DGMM). Accordingly, we propose to adopt the particle filter (PF) in modeling DGMM for saliency detection of face videos, so called PF-DGMM. Finally, the experimental results show that our PF-DGMM approach significantly outperforms other state-of-the-art approaches in saliency detection of face videos.**

***Index Terms*—Face video, visual attention, Gaussian mixture model.**

I. Introduction

DURING the past two decades, saliency detection has become increasingly popular due to its wide application in multimedia processing tasks, such as object segmentation [1]–[5], video quality assessment [6], perceptual video coding [7] and thumbnail generation [8]. Recently, object detection has also taken advantage of visual saliency in segmenting salient objects, called salient object detection [9]–[14]. Visual saliency [15] indicates how much each pixel or region attracts human attention. The first study on visual saliency was performed on images in 1998, when Itti and Koch [16] found that intensity, color and orientation information in an image can be employed to predict image's saliency map. Afterwards, they extended their work to video saliency detection [17]. Recently, a great number of approaches, such as [18]–[29], have been proposed to model saliency in videos. Those saliency-detection approaches are generally driven by biologically-inspired features, which rely heavily on the unmatured study of the human visual system (HVS).

M. Xu , Y. Ren, Z. Wang and J. Liu are with the School of Electronic and Information Engineering, Beihang University, Beijing, 100191 China. X. Tao is with Electronic Engineering Department, Tsinghua University. This work was supported by NSFC under grant numbers 61573037 and 61471022, and Fok Ying-Tong education foundation under grant 151061.

Most recently, data-driven approaches [30]–[42] which learn to bridge the gap between image/video features and saliency, have become prevalent in both video and image saliency detection. These data-driven approaches have found that some high-level features are indeed attractive to visual attention. In particular, face is an obvious high-level feature to attract visual attention, and thus many top-down approaches have incorporated face as a channel for saliency detection of face images [39]–[42]. Specifically, Cerf *et al.* [39] investigated from eye tracking data that face is strongly correlated with visual attention, and they thus proposed to combine face channel with Itti's model [16] for detecting saliency of images including a face. Later, Zhao *et al.* [40] found that the face and orientation channels are usually more important than color and intensity channels. Therefore, they learnt the optimal weights of different channels using least square fitting of eye tracking data, further improving the saliency detection performance of [39]. Most recently, Xu *et al.* [42] proposed to model the saliency distribution of the face region using a Gaussian mixture model (GMM) [43], which is learnt from the training data using the conventional expectation maximization (EM) algorithm. The above approaches consider images with faces, significantly advancing the development of the top-down saliency detection of images.

Face videos [44] are currently undergoing explosion of growth, due to the emerging video conferencing applications, such as FaceTime and Skype. Actually, face also plays an important role in predicting saliency of video conferencing, similar to its important role in saliency detection of face images. Moreover, as we analyze in this paper (Section III-C), face attracts more visual attention in videos (77.7% fixations) than that in image (62.3% fixations). Thus, face is a significant cue for saliency detection in videos. However, most of the existing video saliency detection approaches [17], [18], [21]–[23], [27]–[29] make use of the bottom-up information, such as motion vector, flicker, as well as spatial and temporal correlation, to detect video saliency. Most recently, [45] has been proposed to detect saliency in multiple-face videos by finding which face attracts the most visual attention, but there is no work on modelling distribution of attention within a single face. On the other hand, although videos are composed of images, they are fundamentally viewed differently, because the dynamic changes of pictures in videos can be seen as saliency cues. Thus, video saliency cannot be precisely predicted merely by the assembly of image saliency, as shown in Figure 1. Figure 1 further shows that saliency of face can be modeled as dynamic GMM (DGMM) in videos, in which

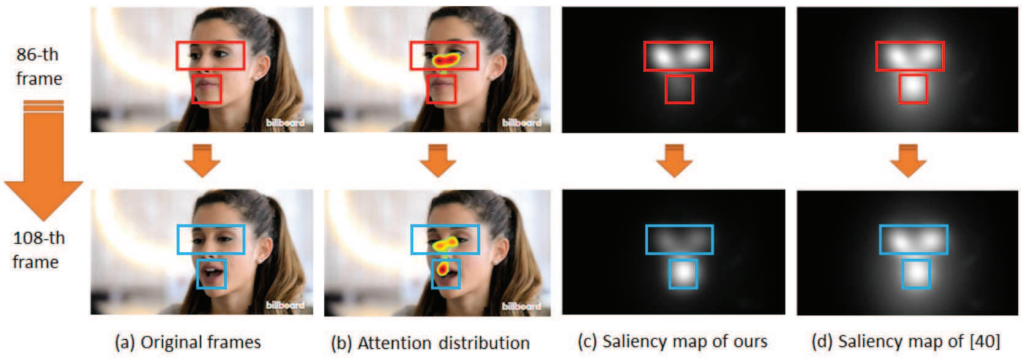


Fig. 1: The comparison of video saliency maps generated by our approach and [42]. Note that [42] is a saliency detection approach for images, while ours work on videos. Here, the saliency maps of [42] are generated by regarding each video frame as a still image. The visual attention distribution by 40 subjects is also shown in this figure. When the mouth moves, visual attention is transferred from eye regions to mouth region.

the GMM distribution of visual attention in face images [42] varies across video frames. It is because visual attention on the mouth increases when the mouth is moving.

In this paper, we establish an eye tracking database for face videos and find that visual attention is attracted by faces in videos. We further find from our database that attention on face prefers to distributing in the regions of facial features, i.e., nose, eyes and mouth. By considering the attention distribution of nose, eyes and mouth as different Gaussian models, GMM is applied to model intra face saliency in each video frame. Beyond, we find that attention distribution on face and facial features vary with respect to face size and mouth movement. Upon these findings, we propose a particle filter based DGMM (called PF-DGMM) approach to detect intra saliency of face videos, modeled by DGMM distribution. In contrast to Xu's static learnt GMM [42] for image saliency detection, the parameters of GMM are dynamically modified according to the mouth movement to obtain the DGMM distribution in our approach. Such dynamic parameters can be learnt from training fixations in our data-driven approach. It is worth pointing out that particle filter (PF) [46] is a sequential model estimation technique that is suitable for monitoring dynamic processes [47]. Thus, for each processed video frame our PF-DGMM approach updates the parameters of GMM, i.e., mean, standard deviation and weights of each Gaussian component, by adopting the PF algorithm. To be more specific, we take the parameters of Gaussian components as particles, and constantly update the weights of the particles to monitor dynamic changes of attention in face videos, according to face size and mouth movement.

The PF-DGMM approach is proposed in this paper for saliency detection in face videos, which is based on the learned GMM distribution of our previous approach [42] for face images. However, our PF-DGMM approach models the saliency distribution of face in videos using dynamic GMM (DGMM), in which the GMM parameters vary across video frames. In contrast, our previous approach [42] only focuses on static GMM distribution of face saliency for images. Towards DGMM, the dynamic changes of GMM parameters are learned in the PF-DGMM approach by designing a PF, whereas the static GMM parameters are estimated in [42] by the EM algorithm. The main contributions of this paper are listed as follows.

- We establish an eye tracking database[1] for face videos, which contains the fixations of 40 subjects viewing 76 single-face videos.
- We analyze the factors that influence the attention distribution on face videos, via mining our established eye tracking database.
- We propose a PF-DGMM approach for learning DGMM distribution as detected saliency of face videos, in light of our analysis on eye tracking database.

There are some potential applications for our saliency detection approach. For example, in compressing video conferencing, the bits can be assigned according to the face saliency detected by our approach, which satisfies the dynamic distribution of human attention within the face region. Consequently, the perceptual quality can be improved in video conferencing applications. Furthermore, in other applications, e.g., encoding or rendering video conferencing, the limited computational resources can be also allocated according to the face saliency detected by our PF-DGMM approach.

II. Literature Review

The existing approaches on saliency detection can be classified into two categories: either heuristic or data-driven models. In the following, we briefly review the video saliency detection literature on these two categories, respectively.

A. Heuristic approaches

Heuristic approaches aim at developing the computational models on image features for saliency detection, according to the understanding of the HVS. During the past decade, a large number of heuristic saliency detection approaches [17]–[21], [24]–[28] have been proposed for video saliency detection. At the beginning, Itti *et al.* [17] found that the dynamic features of motion and flicker contrast are correlated with visual attention. Therefore, [17] combined these two dynamic features with Itti's image saliency model [16] for

[1]The database is available online $https://github.com/RenYun2016$.

detecting saliency in videos. Later, a novel feature called *surprise* was defined in [18] to measure how visual changes attract human attention, based on the Kullback-Leibler (KL) divergence between spatio-temporal posterior and prior beliefs. Given the feature of *surprise*, [18] developed a Bayesian framework for video saliency detection. Besides, some other Bayesian framework related approaches, e.g., [19] and [20], were proposed for video saliency detection. Recently, some advanced video saliency detection approaches [24], [25] have been proposed, also from the biologically-inspired aspect. For example, Lin *et al.* [25] utilized earth mover's distance (EMD) to measure the center-surround difference in spatio-temporal receptive filed, yielding dynamic saliency maps for videos. Hou *et al.* [26] proposed to explore the information divergence model for image saliency detection, and then information divergence is exploited to improve Bayesian surprise for saliency detection in videos. Besides, [27] and [28] analyzed that some compressed domain features (e.g., motion vector) are highly correlated with heuristic saliency detection features (e.g., object motion), and they applied these features to video saliency detection. Recently, saliency detection has been incorporated for detecting object-level saliency [9]–[14], in the field of salient object detection. In particular, [11] proposed a spatiotemporal saliency energy function as a heuristic cue, which encourages the spatiotemporal consistency of video saliency maps, significantly improving the performance of salient object detection. In addition, low rank decomposition was applied in [12], [13] to detect salient objects.

However, the understanding of the HVS is still in its infancy, and heuristic saliency detection thus has a long way to go yet. Recently, machine learning techniques have emerged as a possible way to generalize visual attention model from eye tracking data. They can be seen as data-driven approaches. These data-driven video saliency detection approaches are reviewed in the following.

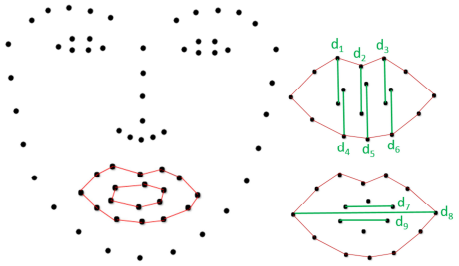
B. Data-driven approaches

The central of data-driven saliency detection approaches is learning visual attention models from eye tracking data, which are obtained using an eye tracker to record fixations of several subjects on displayed images or videos. Accordingly, the existing data-driven saliency detection approaches can be further categorized into static and dynamic saliency prediction tasks. The static task mainly refers to image saliency detection, whereas dynamic tasks primarily concentrate on video saliency detection. In the following, we review the data-driven saliency detection approaches from the aspects of these two tasks.

In static saliency detection, there is a large number of data-driven approaches for generic images [30]–[33], [48]–[52] and specific images [37]–[42]. For detecting saliency of generic images, the representative approach is [30]. In [30], a linear SVM classifier is learnt from eye tracking data (1003 images observed by 15 subjects), in which high-, middle- and low-level features are integrated together in predicting saliency regions. Besides, Hua *et al.* [31] proposed to learn middle-level features, i.e., gists of a scene, as the top-down cue for detecting static saliency of generic images. Instead of the aforementioned hand-tuned features, the latest work of [53] automatically learns hierarchical features for image saliency detection. Most recently, deep neural networks (DNNs) have been widely used in the saliency detection of generic images. Deep gaze I [48] was proposed to apply DNNs to automatically learn effective features for saliency detection. Huang *et al.* [32] have developed the saliency in context (SALICON) method to learn high-level semantic features of objects in saliency detection, on the basis of DNNs. A shallow-structured DNN [33] was explored to further advance image saliency detection. Meanwhile, Saumya *et al.* [49] proposed a new saliency map model formulated by generalized Bernoulli distribution, which is learnt by a DNN architecture. DNNs have also advanced the state-of-the-art salient object detection [50]. Some new DNN architectures [51], [52] were developed to simultaneously deal with saliency detection and salient object detection.

There also exist several works on static saliency detection of some specific images. Focusing on gray images with natural-scene, a gaze-attentive fixation finding engine (GAFFE) [37] was proposed to learn the bandpass filters of both luminance and contrast from eye tracking data, which are used as the low-level features for saliency detection. Afterwards, dictionary learning was applied in [38], together with sparse coding, to learn the patterns of salient and non-salient regions for gray images with nature scene. Beyond, more approaches [39]–[42] work on predicting saliency of images with face, since face is an obvious cue for drawing visual attention. For example, Cerf *et al.* [39] found out that face is an important top-down feature to receive attention, as the faces were fixed on in 88.9% within two fixations (7 subjects viewing 150 images with face) in their eye tracking experiment. Therefore, they proposed to combine Viola-Jones (VJ) face detector [54] with Itti's model [16] for saliency detection over images with face. Afterwards, [40] was proposed to learn the weights of top-down features (i.e., face) using least square fitting of eye tracking data, thus improving the saliency detection performance of [39]. Later, Jiang *et al.* [41] developed some high-level features that are related to human faces, to predict saliency in a scene with multiple faces. Those high-level features include face size, pose and location. For single-face image, [42] has been proposed to precisely model saliency of face region, via learning the fixation distributions of face and facial regions.

In dynamic saliency detection, most existing data-driven approaches [34]–[36], [55]–[60] concentrate on generic videos. In particular, Rudoy *et al.* [34] proposed a novel method to predict the dynamic saliency of generic videos, which learns the conditional saliency map from fixations across several consecutive video frames. As a result, the inter-frame correlation of visual attention is taken into account for video saliency detection. In [55], Zhao *et al.* proposed a fixation bank approach for video saliency detection, in which a bank is built from the primitive low-level features of color, intensity, orientation and motion. Recently, both spatio-temporal coherency and low rank analysis have been applied [35] for locating salient motion in videos. In [56], dynamic adaptive whitening saliency (AWS-D) was proposed for detecting video saliency, which reduces the perceptual redundancy in locating



(a) 66-point PDM (b) Mouth

Fig. 2: An example of PDM for the face and mouth.

salient pixels based on the second-order statistics. The most recent work of [57] studied how to apply learning algorithms in effectively integrating low-level and high-level features, for video saliency detection. Rather than free-view saliency detection, a probabilistic multi-task learning method was developed in [36] for the task-driven video saliency detection, in which the "stimulus-saliency" functions are learnt from the eye tracking data as the top-down attention models. In [58], Mathe and Sminchisescu established a new eye tracking database and then proposed saliency detection models, for the task of action recognition rather than free-view. In [59], Mauthner *et al.* proposed an encoding based video saliency detection approach for the task of activity recognition. Different from the above free-view and task-driven saliency detection approaches, [60] proposed a space-time saliency approach, which locates salient frames, rather than pixels or objects. However, the above approaches were designed for generic saliency detection of videos. They are thus ineffective in saliency detection of the specific face videos, since they do not explore dynamic variation of attention distribution within face regions.

For specific videos, [45] extended the work of [41] to detect saliency in multiple-face videos by proposing multiple hidden Markov model (M-HMM). In [45], M-HMM is used to predict the possibility of each face on attracting visual attention. Although the face has been considered as the high-level feature for saliency detection of generic videos, few works aim at modelling the saliency distribution within face region for face videos. In fact, the saliency distribution of face is dynamic, which may be influenced by actions of face like mouth movement. Thereby, this paper establishes an eye tracking database for face videos, and then we learn from our database, to predict dynamic distribution of face saliency.

III. Database and Analysis

A. Database

To our best knowledge, there exists no eye tracking database on face videos. Therefore, we conducted the eye tracking experiment to obtain an eye tracking database for several videos containing faces. The database is composed of 76 face videos selected from the 300-VW [61] database and YouTube. Among them, 71 videos contain one face, and the rest have two faces. The resolutions of all 76 videos in our database are 1280×720, and their frame rates are around 30Hz. There are 40 subjects[2] involved in the experiment to watch all 76 videos, including 24 males and 16 females aging from 21 to 35.

During the experiment, a 23-inch 1080p LCD screen, integrated in the eye tracker, was used to display the videos. The videos were displayed at their original resolution (720p), and their display order was random to reduce the eye fatigue effect on the eye tracking results. All 40 subjects were asked to watch these video without any task. Besides, the fixations of those 40 subjects on each video were recorded by a Tobii X2-60 eye tracker at the sampling rate of 60Hz.

Finally, $1,119,368$ fixations over $30,936$ frames of 76 videos were collected in our database. Our database can be freely downloadable via https://github.com/RenYun2016/PF-DGMM, for facilitating the future research.

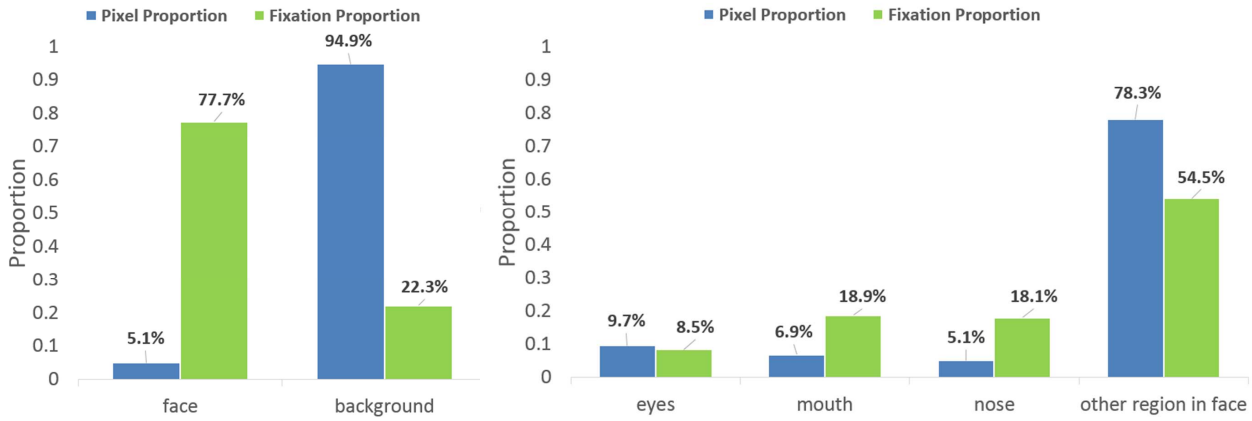
B. Features extraction

It is intuitive that face related features significantly influence the distribution of visual attention on face videos. Before analyzing the relationship between face and attention distribution, this section addresses the extraction of face-related features in videos, including face, facial features and mouth movement.

Face and Facial Features. First of all, it is necessary to extract face and facial features for attention analysis and saliency detection. In this paper, we follow the way of [42] to automatically segment the regions of the face and facial features, by leveraging the face alignment algorithm [62]. Specifically, 66 landmark points are located according to point distribution model (PDM) [62]. Then, some landmark points are connected to precisely obtain the contours of face and facial features. Upon the contours, the regions of face and facial features can be extracted. Figure 2-(a) shows an example for the extraction of the face and facial features, based on the 66-point PDM.

Intensity of Mouth Movement. We empirically find that the distribution of visual attention on the face in a video is correlated with the movement intensity of the mouth. Therefore we need to measure the intensity of mouth movement. In this paper, we use the following way to quantify the mouth movement intensity of the t-th frame (denoted by D_t) in a face video. As we show in Figure 2-(b), the mouth region is composed of 18 landmark points. Generally speaking, D_t can be determined by the difference of the width or height across neighboring frames, which may be computed upon the 18 detected points of the mouth region. Since the shape of mouth movement varies, both width and height are included in calculating D_t, corresponding to variation in the horizontal and vertical directions, respectively. However, there may be some errors in detecting the 18 points of mouth. To reduce the impact of these errors on D_t, D_t is calculated by averaging over more than one Euclidean distance, alongside the horizonal or vertical direction. In our approach, we compute on 9 Euclidean distances, i.e., d_1, d_2, d_3, d_4, d_5 and d_6 for the vertical distance, and d_7, d_8, d_9 for the horizontal distance. Refer to Figure 2-(b) for more details about these 9 Euclidean distances for computing mouth movement intensity. Finally,

[2]All 40 subjects have either corrected or uncorrected normal eyesight.



(a) Face and background (b) Subregions in face

Fig. 3: Proportions of fixations and pixel numbers in different regions,counted on all 76 videos in our database. (a) shows the proportions for the regions of face and background, and (b) illustrates the proportions of fixations over different regions of face.

the intensity of mouth movement at the t-th frame can be calculated as follows,

$$D_t = \sum_{i=1}^{9} \frac{|d_i^{(t)} - d_i^{(t-t')}|}{\min(d_i^{(t)}, d_i^{(t-t')})}, \quad (1)$$

where $d_i^{(t)}$ and $d_i^{(t-t')}$ are the i-th distance of mouth at the t-th and $(t-t')$-th frames, respectively. In (1), the absolute difference between $d_i^{(t)}$ and $d_i^{(t-t')}$ is divided by their minimal value to model the relative intensity of mouth movement. Here, t' is the number of frames used to determine mouth variation, which is related to the frame rate of the video. According to the theory of persistence of vision [63], there exists approximately 0.1 second residual for motion perception. Since the interval between the t-th and $(t-t')$-th frames needs to be larger than motion perception in (1), t' is computed by

$$t' = \text{round}(0.1 \cdot fr). \quad (2)$$

where fr is the frame rate of a video. Finally, D_t can be obtained using (1) and (2). Obviously, a large value of D_t means the high intensity of the mouth movement, which may increase visual attention on mouth region.

C. Database analysis

We investigate the intrinsic factors which have an impact on visual attention to face videos, by analyzing the fixations obtained from the 76 videos in our database. Intuitively, visual attention is not uniformly distributed in the face region of face videos. Upon the extraction of face related features, we analyzed the attention distribution within face. Then, we have the following observations. Note that the technique on extracting face and facial features for our database analysis is to be discussed in Section III-B.

Observation 1: For a video, face attracts significantly more visual attention than background, and within face region, facial features (i.e., eyes, nose and mouth) are more salient than other regions of the face.

First, we show in Figure 3-(a) the proportions of fixations and pixels belonging to face and background, respectively, for all 76 videos. As seen in Figure 3-(a), although the face region only takes up 5.1% pixels in video frames, it attracts 77.7% visual attention. Compared to 62.3% fixations attracted by face[3] in images [42], face region is more salient in drawing visual attention in videos. Besides, Figure 3-(b) illustrates the proportions of pixels and fixations within face region. We can see from this figure that facial features consume 21.7% pixels in face region (9.7% for two eyes, 6.9% for mouth and 5.1% for nose), whereas they draw 45.5% fixations (8.5% for eyes, 18.9% for mouth and 18.1% for nose). Thus, we can conclude that facial features are more salient than other regions in face for a video. This completes the analysis of Observation 1.

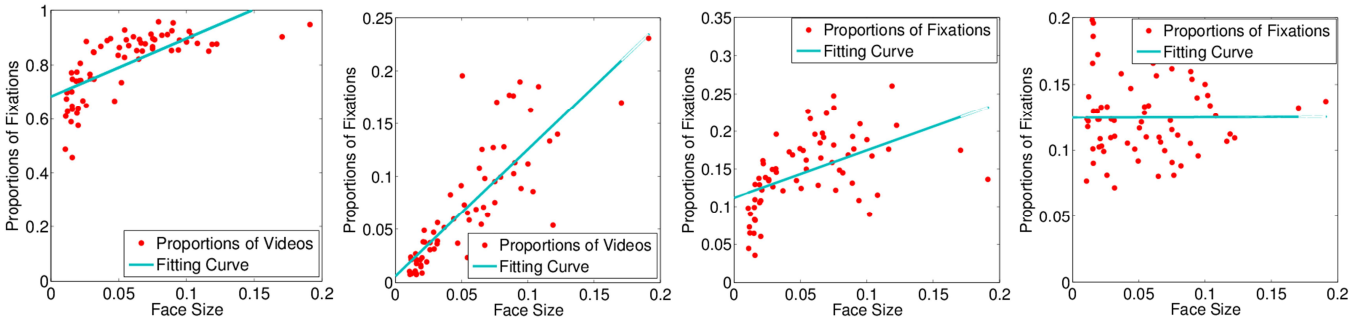
Observation 2: Visual attention on the face, eyes and nose increases along with the enlarged size of face in a video, whereas the attention on mouth is invariant to the face size.

Figure 4 shows the proportions of fixations belonging to the regions of the face and facial features for different face sizes in the 76 videos of our database. In this paper, we define face size as the proportion of pixels belonging to the face region in a video frame. The face region is segmented using the way of Section III-B. In this paper, we define face size as the proportion of pixels belonging to the face region in a video frame. The face region is segmented using the method described in Section III-B. In Figure 4, the fitting curves are plotted to reflect the general trend that how proportions of fixations in facial features change alongside increased face sizes. From Figure 4, we can find out that when the face size becomes larger, the proportions of fixations in face, eyes and nose increase. However, the proportions of fixations in mouth are almost unchanged, implying that visual attention on mouth is invariant to the size of face in the video. This completes the analysis of Observation 2.

Observation 3: Visual attention on eyes is not affected by the eye blink, whereas more attention is drawn by the mouth when it is moving.

Before figuring out the relationship between visual attention and mouth movement or eye blink, we obtained the ground-

[3]Note that face averagely has 5.7% pixels in the whole image.



(a) Face (b) Eyes (c) Nose (d) Mouth

Fig. 4: Proportions of fixations on face and facial features versus face sizes, for all 76 videos of our database. Each dot in the figure stands for the statistical result of one video. The least square fitting curves of linear regression on fixation proportions of all frames in 76 videos are provided (blue lines). The Spearman rank correlation coefficients between face size and fixation proportions in each region are (a) 0.82, (b) 0.88, (c) 0.65 and (d) 0.09.

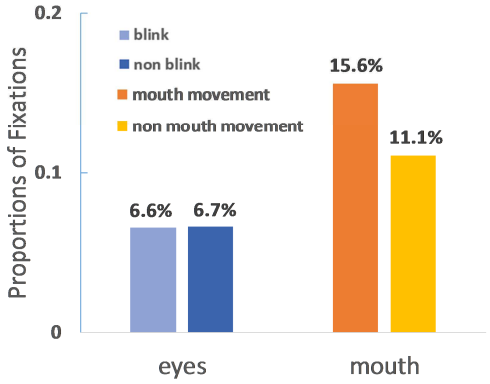


Fig. 5: Proportions of fixations in eyes and mouth with and without movement per video frame.

truth eye blink and mouth movement, by manually annotating all 76 videos of our database[4]. Then, the statistical results of fixations versus mouth movement and eye blink are shown in Figure 5, for all 76 videos of our eye tracking database. From this figure, we can find that the proportions of fixations on eyes are almost the same, whether the eyes blink or not. This implies that visual attention on eyes is invariant to eyes movement. On the contrary, more fixations are drawn by mouth (from 11.1% to 15.6%) when mouth movement occurs. This completes the analysis of Observation 3.

Observation 4: Visual attention on mouth increases along with the enlarged intensity of mouth movement.

Figure 6 plots the fixation proportions of the mouth at different intensities for mouth movement, with the scatter analysis of all fixations on mouth regions from our database. Here, the intensity of mouth movement is measured using the method described in Section III-B. We can see from Figure 6 that attention on mouth generally increases with the increased intensity of mouth movement, the Spearman rank correlation coefficient of which (=0.24) is much larger than that between face size and fixation proportion in mouth regions (=0.09). This completes the analysis of Observation 4.

[4]Three volunteers annotated the movements of the eyes and mouth in all frames of the 76 videos. Then, the ground-truth annotations of the eye and mouth movements were obtained by majority voting. These annotations are also provided along with our eye tracking database.

IV. The Proposed Approach

A. Framework

In this section, we mainly discuss the proposed approach for detecting saliency of face videos. In our approach, we follow the basic way of [39] and [42] to predict the saliency map of each video frame by

$$\mathbf{M} = w_C \mathbf{S}_t^C + w_I \mathbf{S}_t^I + w_O \mathbf{S}_t^O + w_F \mathbf{S}_t^F. \quad (3)$$

In (3), $\mathbf{S}_t^C$, $\mathbf{S}_t^I$, $\mathbf{S}_t^O$ and $\mathbf{S}_t^F$ are the saliency maps of the feature channels of color, intensity, orientation and face at the t-th video frame. w_C, w_I, w_O and w_F are the weights corresponding to each feature channel. In this paper, they are learnt by the least square fitting on the training data.

Our approach adopts Itti's model [16] to yield saliency maps $\mathbf{S}_t^C$, $\mathbf{S}_t^I$ and $\mathbf{S}_t^O$ for the channels of low-level features. In addition to the low-level features, the face is incorporated in our approach. This satisfies Observation 1 that face receives most visual attention in a video. In the following, we aim at computing the saliency distribution $\mathbf{S}_t^F$ of the face channel, for saliency detection of face videos. Observations 2 and 4 have verified that the variation of face size and the action of mouth movement in a video dynamically change the distribution of visual attention on face. Thus, our approach models the dynamic visual attention on face regions by proposing a new distribution model called DGMM. It is worth pointing out that it is different from the previous works of [39] and [42], which only use static GM or GMM to model visual attention on the face.

The overall framework of our saliency detection approach is shown in Figure 7. Next, we introduce DGMM to model the dynamic distribution of visual attention on the face for videos.

B. DGMM for single-frame saliency detection

When processing the t-th frame of a face video, we can model its saliency distribution $\hat{\mathbf{S}}_t^F$ as follows,

$$\hat{\mathbf{S}}_t^F = \sum_{i=1}^{5} \hat{\pi}_t^i \mathcal{G}^i = \sum_{i=1}^{5} \hat{\pi}_t^i \mathcal{N}(\hat{\mu}_t^i, \hat{\sigma}_t^i), \quad (4)$$

where $\mathcal{N}(\cdot)$ is the Gaussian distribution, and $\hat{\pi}_t^i$, $\hat{\mu}_t^i$ and $\hat{\sigma}_t^i$ are the weight, mean and standard deviation of the i-th GM

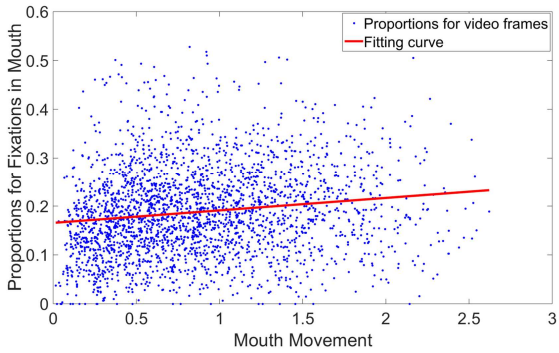


Fig. 6: Proportions of fixations in mouth at different intensities for mouth movement. The Spearman rank correlation coefficient here is 0.24.

$\mathcal{G}^i$. Similar to [42], $\mathcal{G}^1$, $\mathcal{G}^2$, $\mathcal{G}^3$, $\mathcal{G}^4$ and $\mathcal{G}^5$ correspond to the GMs of face, left eye, right eye, nose and mouth. We use the same method as [42] to learn means and standard deviations of these GMs.

In the following, we propose to learn weights $\{\hat{\pi}_t^i\}_{i=1}^5$ of each GM from training data, in accord with our observations of Section III-C. Assume that we have L training video frames, and their corresponding fixation maps are $\{\mathbf{S}_l^*\}_{l=1}^L$. Note that the fixation maps are obtained by applying a 2D Gaussian filter to smooth the ground-truth fixations in each training frame. Then, GM weights $\{\hat{\pi}_l^{i*}\}_{i=1}^5$ of each training frame can be obtained by solving the following ℓ_2-norm optimization formula:

$$\min_{\{\hat{\pi}_l^{i*}\}_{i=1}^5} ||\sum_{i=1}^5 \hat{\pi}_l^{i*}\mathcal{G}_l^{i*} - \mathbf{S}_l^*||_2 \quad \text{s.t.} \quad \sum_{i=1}^5 \hat{\pi}_l^{i*} = 1, \hat{\pi}_l^{i*} > 0, \tag{5}$$

where $\{\mathcal{G}_l^{i*}\}_{i=1}^5$ are the GMs of the l-th training frame. In our approach, we utilize the CVX[5] [64] to solve the above formula.

Given $\{\hat{\pi}_l^{i*}\}_{i=1}^5$ of all training frames, we can learn $\{\hat{\pi}_t^i\}_{i=1}^5$ as follows. Observation 2 showed that visual attention to the face, eyes and nose depends on face size. Therefore, we learn the relationship between face size and $\{\hat{\pi}_t^i\}_{i=1}^4$, using polynomial regression over the training data $\{\hat{\pi}_l^{i*}\}_{i=1}^4$ and s_l^*, where $\{s_l^*\}_{l=1}^L$ are the face sizes of training frames. Then, $\{\hat{\pi}_t^i\}_{i=1}^4$ can be represented by

$$\hat{\pi}_t^i = \sum_{k=0}^{K_1} a_k^i \cdot (s_t)^k, \quad i = 1, 2, 3, 4 \tag{6}$$

where s_t is the size of face at the t-th test video frame. In (6), $\{a_k^i\}_{k=0}^{K_1}$ are the polynomial parameters to be learnt, with K_1 being the order of the polynomial function. In Section V-A, we provide more details about the values of K_1 and $\{a_k^i\}_{k=0}^{K_1}$, which are obtained by training in our experiments.

Now, we move to the computation of $\hat{\pi}_t^5$, which is the weight of the GM belonging to the mouth. Since Observation 4 indicated that the saliency of mouth in a video is correlated with the intensity of mouth movement, we also use the polynomial regression to learn $\hat{\pi}_t^5$. Recall that D_t is the intensity of mouth movement at the t-th frame, as denoted in Section III-B. Then, we have

$$\hat{\pi}_t^5 = \sum_{k=0}^{K_2} b_k \cdot (D_t)^k. \tag{7}$$

In (7), $\{b_k\}_{k=0}^{K_2}$ and K_2 are the polynomial parameters and the order of the regression, to be discussed in Section V-A with more details.

Finally, face saliency $\hat{\mathbf{S}}_t^F$ of each single frame can be generated, based on (4). However, it is clear that a sudden change in the saliency distribution of the face across neighboring frames is rare. In other words, the saliency map of a video frame not only depends on its observed features, but also on the saliency distribution of face at the previous frames. Due to this, we propose the PF-DGMM algorithm in the next subsection, which applies PF to smooth DGMM saliency distribution across frames. As such, both the past saliency and the observed features can be considered in a uniform framework to output the final saliency map of the dynamic face channel $\mathbf{S}_t^F$ at frame t.

C. PF-DGMM for video saliency detection

To satisfy temporal consistency in a video, we need to enable the smooth transition of dynamic face saliency across frames. In other words, the DGMM saliency distribution of each frame should depend on the observed features at the current frame as well as the estimated DGMM of previous frames. The PF estimates the internal states in dynamic systems given sequential observations, which has been widely used to make inferences on sequential data. Thus, the PF is integrated with the DGMM distribution, so called the PF-DGMM algorithm, for modeling the dynamic face saliency of videos.

Mathematically, $\mathbf{S}_t^F$, which is the final saliency distribution of face at the t-th frame, should be predicted on the basis of $\mathbf{S}_{t-1}^F$ and $\hat{\mathbf{S}}_t^F$. Note that both $\mathbf{S}_t^F$ and $\hat{\mathbf{S}}_t^F$ are modeled by GMM distributions in each video frame. In our PF-DGMM algorithm, we need to track the dynamic changes in videos, to adjust the values of $\{\hat{\pi}_t^i\}_{i=1}^5$, $\{\hat{\mu}_t^i\}_{i=1}^5$ and $\{\hat{\sigma}_t^i\}_{i=1}^5$ of GMM by (4). After the adjustment, $\{\pi_t^i\}_{i=1}^5$, $\{\mu_t^i\}_{i=1}^5$ and $\{\sigma_t^i\}_{i=1}^5$ can be yielded as the parameters of $\mathbf{S}_t^F$, for the final output of face saliency in a video. Obviously, the centroid μ_t^i of each GM is in the center of face or facial features, and we therefore

[5]CVX is a Matlab-based toolbox for solving the problem of convex optimization.

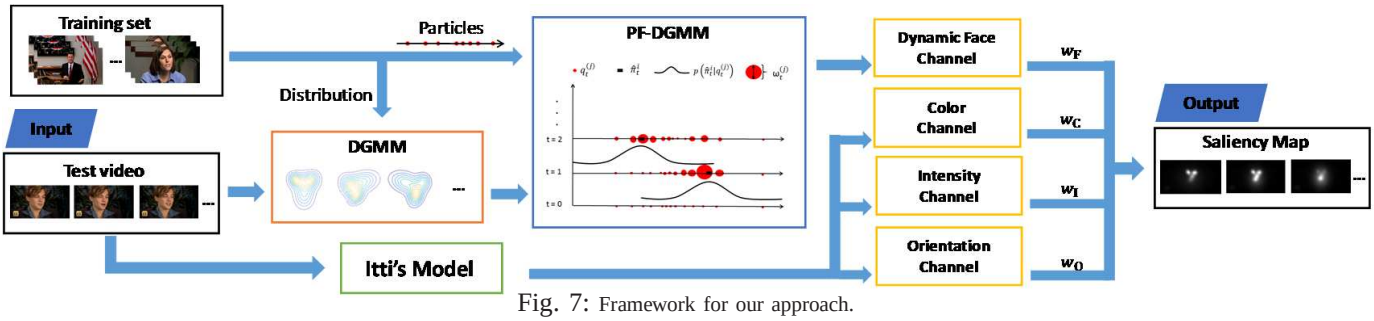


Fig. 7: Framework for our approach.

Algorithm 1 Summary of our PF-DGMM algorithm

Input: Means of DGMM $\{\mu^i_{1:t}\}^5_{i=1}$ and DGMM parameters of $\hat{\mathbf{S}}^F_{1:t}$: $\{\hat{\pi}^i_{1:t}\}^5_{i=1}$ and $\{\hat{\sigma}^i_{1:t}\}^5_{i=1}$, which are obtained in Section V-B.
Output: Saliency map of the dynamic face channel $\mathbf{S}^F_t$.
Initialize: The total number of frames: T; the total number of particles: J; the initial sets of training parameters: $\{\hat{\pi}^{i*}_l\}^5_{i=1}$; $\{\hat{\sigma}^{i*}_l\}^5_{i=1}$ $(l = 1, \ldots, L)$; and the weights of particles: $\omega^{(j)}_{\pi^i_0} = \frac{1}{J}$ and $\omega^{(j)}_{\sigma^i_0} = \frac{1}{J}$.
for $t = 1$, $t \leq T$, $t++$ **do**
  for $i = 1$, $i \leq 5$, $i++$ **do**
    Sample particles of parameters π^i_t and σ^i_t from training set: $q^{(j)}_{\pi^i_t} \in \{\hat{\pi}^{i*}_l\}^L_{l=1}$ and $q^{(j)}_{\sigma^i_t} \in \{\hat{\sigma}^{i*}_l\}^L_{l=1}$.
    Estimate the likelihood $p(\hat{\pi}^i_t|q^{(j)}_{\pi^i_t})$ and $p(\hat{\sigma}^i_t|q^{(j)}_{\sigma^i_t})$ with (14), given $\hat{\pi}^i_t$ and $\hat{\sigma}^i_t$.
    Update weights $\omega^{(j)}_{\pi^i_t}$ and $\omega^{(j)}_{\sigma^i_t}$ using (13) with previous weights $\omega^{(j)}_{\pi^i_{t-1}}$ and $\omega^{(j)}_{\sigma^i_{t-1}}$.
    Calculate $\mathbb{E}(\pi^i_t)$ and $\mathbb{E}(\sigma^i_t)$ using (15).
    Normalize weights $\omega^{(j)}_{\pi^i_t}$ and $\omega^{(j)}_{\sigma^i_t}$ via (16).
    Set $\pi^i_t = \mathbb{E}(\pi^i_t)$ and $\sigma^i_t = \mathbb{E}(\sigma^i_t)$.
  end for
  Calculate saliency distribution over face $\mathbf{S}^F_t$ by $\sum^5_{i=1} \pi^i_t \mathcal{N}(\mu^i_t, \sigma^i_t)$, where $\mathcal{N}(\cdot)$ is the Gaussian distribution.
  Return $\mathbf{S}^F_t$.
end for

use detected face and facial features to obtain $\{\mu^i_t\}^5_{i=1}$. For $\{\pi^i_t\}^5_{i=1}$ and $\{\sigma^i_t\}^5_{i=1}$, we can track the dynamic change based on the PF as follows.

Following Bayesian tracking theory [47], we take $\{\pi^i_t\}^5_{i=1}$ and $\{\sigma^i_t\}^5_{i=1}$ as random variables. These variables can be predicted by

$$p(\{\pi^i_t\}^5_{i=1}, \{\sigma^i_t\}^5_{i=1}|\{\hat{\pi}^i_{1:t}\}^5_{i=1}, \{\hat{\sigma}^i_{1:t}\}^5_{i=1}), \tag{8}$$

in which the distribution of $\mathbf{S}^F_{t-1}$ is embedded in $\{\hat{\pi}^i_{1:t-1}\}^5_{i=1}$ and $\{\hat{\sigma}^i_{1:t-1}\}^5_{i=1}$. Besides, $\{\hat{\pi}^i_t\}^5_{i=1}$ and $\{\hat{\sigma}^i_t\}^5_{i=1}$ encode observed $\hat{\mathbf{S}}^F_t$ of the current frame. Consequently, the values of $\{\pi^i_t\}^5_{i=1}$ and $\{\sigma^i_t\}^5_{i=1}$ depend on $\mathbf{S}^F_{t-1}$ and $\hat{\mathbf{S}}^F_t$. It is obvious that $\{\pi^i_t\}^5_{i=1}$ and $\{\sigma^i_t\}^5_{i=1}$ in DGMM are independent of each other. Thereby, (8) can be decomposed as

$$p(\{\pi^i_t\}^5_{i=1}|\{\hat{\pi}^i_{1:t}\}^5_{i=1}), \tag{9}$$

and

$$p(\{\sigma^i_t\}^5_{i=1}|\{\hat{\sigma}^i_{1:t}\}^5_{i=1}). \tag{10}$$

Assuming that the elements of $\{\pi^i_t\}^5_{i=1}$ are not correlated with each other in (9), we can utilize the following expectation of π^i_t:

$$\mathbb{E}(\pi^i_t) = \int p(\pi^i_t|\hat{\pi}^i_{1:t})\pi^i_t d\pi^i_t, \tag{11}$$

to estimate π^i_t. However, the probability distribution of $p(\pi^i_t|\hat{\pi}^i_{1:t})$ in the above equation is not available. Therefore, our PF-DGMM algorithm uses the Monte Carlo method to obtain $p(\pi^i_t|\hat{\pi}^i_{1:t})$ by introducing particles. Note that particles $q^{(j)}_t$ $(j = 1, 2, ..., J)$ are regarded as possible values of π^i_t at the t-th frame, which are initialized to be $\hat{\pi}^{i*}_l$ of the training data. As a result, $p(\pi^i_t|\hat{\pi}^i_{1:t})$ can be calculated by

$$p(\pi^i_t|\hat{\pi}^i_{1:t}) = \sum_{j=1}^{J} \omega^{(j)}_t \delta(\pi^i_t - q^{(j)}_t), \tag{12}$$

where $\delta(\cdot)$ is the Dirac delta function, and $\omega^{(j)}_t$ is the weight of the j-th particle $q^{(j)}_t$. According to PF theory [65], $\omega^{(j)}_t$ can be updated from $\omega^{(j)}_{t-1}$ in the form of

$$\omega^{(j)}_t = \omega^{(j)}_{t-1} p(\hat{\pi}^i_t|q^{(j)}_t). \tag{13}$$

In (13), $p(\hat{\pi}^i_t|q^{(j)}_t)$ is the likelihood of $q^{(j)}_t$ given $\hat{\pi}^i_t$. Consequently, (13) reflects the previous saliency distribution $\mathbf{S}^F_{t-1}$ at the previous frame (encoded in $\omega^{(j)}_{t-1}$), and the observed saliency distribution $\hat{\mathbf{S}}^F_t$ at the current frame (encoded in $p(\hat{\pi}^i_t|q^{(j)}_t)$). Generally speaking, the probability of $p(\hat{\pi}^i_t|q^{(j)}_t)$ increases when $q^{(j)}_t$ approaching to $\hat{\pi}^i_t$. Therefore, $p(\hat{\pi}^i_t|q^{(j)}_t)$ is modeled in our PF-DGMM algorithm by

$$p(\hat{\pi}^i_t|q^{(j)}_t) = \exp\left(|\hat{\pi}^i_t - q^{(j)}_t| - c\right), \tag{14}$$

In the above equation, c is a constant set to 1, to fix the range of probability $p(\hat{\pi}^i_t|q^{(j)}_t)$ to $[0, 1]$. Specifically, the weights of Gaussian components $\hat{\pi}^i_t$ and particles $q^{(j)}_t$ are both in the range of $[0, 1]$. Thus, we have $0 \leq |\hat{\pi}^i_t - q^{(j)}_t| \leq 1$. When $c = 1$, the range of $\exp(|\hat{\pi}^i_t - q^{(j)}_t| - c)$ is $[e^{-1}, 1]$, which is $\subset [0, 1]$. Consequently, probability $p(\hat{\pi}^i_t|q^{(j)}_t)$ is restricted to $[0, 1]$ in the above equation via setting $c = 1$. Given (11) and (12), π^i_t can be obtained by setting it equivalent to its expectation:

$$\mathbb{E}(\pi^i_t) = \sum_{j=1}^{J} \omega^{(j)}_t q^{(j)}_t, \tag{15}$$

with $\omega^{(j)}_t$ obtained from (13) and (14).

At last, we need to normalize $\omega^{(j)}_t$ for the next frame by the following equation:

$$\omega^{(j)}_t = \frac{\omega^{(j)}_t}{\sum_{j=1}^{J} \omega^{(j)}_t}. \tag{16}$$

TABLE I: The learnt polynomial coefficients a_k^i

features	face	left eye	right eye	nose
a_k^i	$i=1$	$i=2$	$i=3$	$i=4$
$k=0$	0.52	-0.07	-0.07	0.11
$k=1$	-7.13	5.49	5.49	7.39
$k=2$	23.09	-17.99	-17.99	-28.01

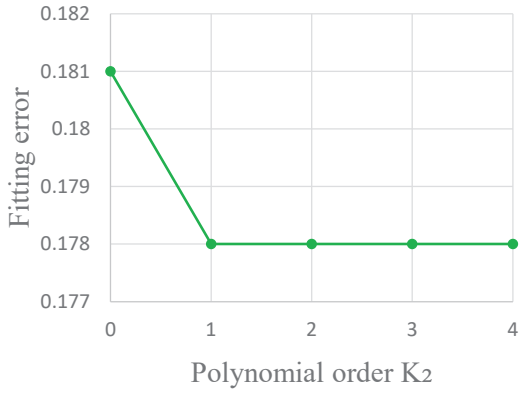


Fig. 8: Fitting error along with different values of polynomial order K_2.

Furthermore, we use the similar way to compute $\{\sigma_t^i\}_{i=1}^5$. Finally, $\mathbf{S}_t^F$ can be achieved as saliency distribution of the face channel at each frame. Our PF-DGMM algorithm is summarized in Algorithm 1.

V. Experimental results

In this section, we evaluate the effectiveness of our PF-DGMM approach in detecting saliency of face videos, via comparing with other 10 state-of-the-art approaches. Next, we present the parameter settings used in our experiment.

A. Settings

In our experiment, all 76 videos in our database, which are discussed in Section III-A, are divided into two groups: the training and test sets. Specifically, the training set is composed of 30 videos randomly selected from our database. It includes 14,592 frames with 523,929 fixations. The other 46 videos form the test set, the results of which are to be reported in Section V-B. Here, all fixations falling into face regions[6] (in total $407,093$ fixations) are selected from 14,592 training frames. Given those training fixations, we apply the following way to learn the weights of GMs.

Learn face-related GM weights. As mentioned in Section IV-B, the weights $\{\hat{\pi}_t^i\}_{i=1}^4$ of each GM, i.e., face, left eye, right eye and nose, should be related to the face size s_t. To this end, we first convolute all fixations in each training frame, to obtain the ground-truth saliency maps $\{\mathbf{S}_l^*\}_{l=1}^L$. Afterwards, the optimal weight of each GM can be estimated for each training frame, i.e., $\{\hat{\pi}_l^{i*}\}_{i=1}^5$, via making DGMM distribution as close to $\{\mathbf{S}_l^*\}_{l=1}^L$ as possible. Given $\{\hat{\pi}_l^{i*}\}_{i=1}^5$ of all 14,592 training frames, the polynomial coefficients $\{a_k^i\}_{k=0}^{K_1}$ can be learnt for encoding the relationship between $\{\hat{\pi}_t^i\}_{i=1}^4$ and s_t with (6). Besides, the order of K_1 in (6) is empirically set to be 2. The learnt coefficients are reported in Table I.

Learn mouth-related GM weights. As shown in (7), the weight of the GM belonging to the mouth, i.e., $\hat{\pi}_t^5$, is correlated with the mouth movement intensity D_t. Similar to learning $\{\hat{\pi}_t^i\}_{i=1}^4$, we train polynomial coefficients $\{b_k\}_{k=0}^{K_2}$ for $\hat{\pi}_t^5$. To determine the order K_2 of the regression in (7), Figure 8 plots the fitting error over all training samples, at different K_2. As shown in this figure, the fitting error is converged when $K_2 \geq 1$. This indicates that the weight of mouth $\hat{\pi}_t^5$ is linearly correlated with D_t. Therefore, we set $K_2 = 1$ in our experiments. Then, $\hat{\pi}_t^5 = b_0 + b_1 \cdot D_t$ achieves the least square error over all training samples, when $b_0 = 0.18$ and $b_1 = 0.02$. Consequently, after the least square fitting, the values of b_0 and b_1 are 0.18 and 0.02 in our experiments, respectively.

Particles for PF-DGMM. Given above learnt $\{a_k^i\}_{i=1}^4$ and b_k, $\{\hat{\pi}_t^i\}_i^5$ can be estimated using (6) and (7) with respect to s_t and D_t, for each training frame. Then, the estimated $\{\hat{\pi}_t^i\}_i^5$ of all training frames are used as the particles of the GM weights. Similarly, we use the estimated standard deviations of all training frames as the particles of standard deviations of GMs. Note that the total number of particles is $14,592$, equaling to the number of all training frames.

B. Evaluation on our database

Evaluation metrics. In our experiment, we utilize the following metrics to evaluate the accuracy of saliency detection: normalized scanpath saliency (NSS), linear correlation (CC) and area under ROC curve (AUC). NSS quantifies the degree of correspondence between human fixation locations and saliency maps. CC measures the strength of the linear correlation between human fixation maps and predicted saliency maps. AUC computes the area under ROC curve, reflecting the tradeoff between the false positive rate and true positive rate. The larger NSS and CC indicate higher accuracy of saliency detection. Besides, the closer the AUC value is to 1, the more accurately the approach can predict human attention.

Evaluation results. Here, we evaluate on all 46 test videos of our database, with other 30 videos being training data. For the evaluation, we compare our approach with 10 other approaches (i.e., Cerf *et al.* [39], Zhou et al. [60], Guo *et al.* [21], Zhao *et al.* [40], Rudoy *et al.* [34], Xu *et al.* [42], Hossein *et al.* [28], WangSeg [14], WangSal [11] and SALICON [32]) to verify the effectiveness of our approach. The comparison results are presented in Table II, in terms of NSS, CC and AUC values with means and standard deviations. As we can see from this table, our approach outperforms other 10 approaches in terms of all three metrics, with 5.20 in NSS, 0.84 in CC and 0.94 in AUC. Specifically, our approach has an improvement of at least 0.41 in NSS and 0.07 in CC compared to other approaches. It is worth pointing out that NSS and CC, especially NSS, are more reasonable metrics than AUC in evaluating saliency accuracy, according to the analysis of [66]. Thus, the results of Table II imply that our approach significantly advances state-of-the-art saliency detection in face videos. Besides, the gain of our approach

[6]All faces across different videos are re-scaled into a uniform coordinate before learning the distribution.

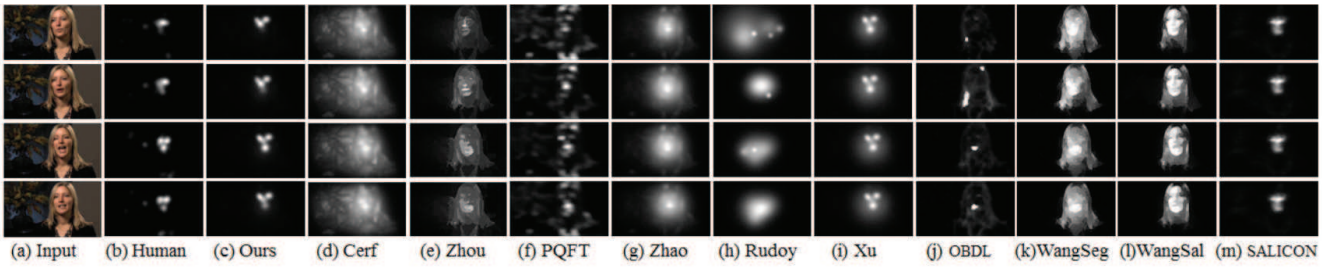


Fig. 9: Saliency maps across different frames (the 84th, 109th, 198th, and 203rd frames) of a randomly selected video, generated by our and other 10 approaches, as well as the human fixations.

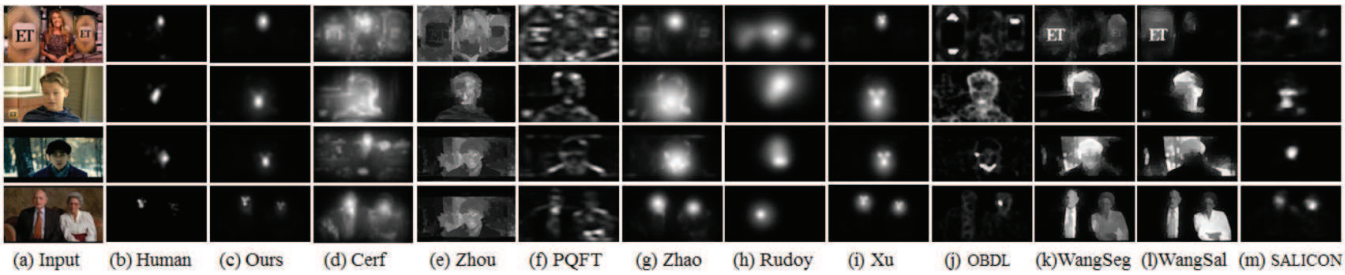


Fig. 10: Saliency maps of some videos at different face sizes, generated by our and other 10 approaches, as well as the human fixations.

over [42] verifies the effectiveness of making GMM dynamic in videos, since face saliency is modeled by GMM in [42] while it is represented by DGMM in our approach.

Subjective results. Figures 9 and 10 illustrate the saliency maps of some selected video frames, generated by our method and other 10 approaches. As we can see from these two figures, the saliency maps of ours are much closer to the ground-truth of human attention maps, than those of other 10 approaches. Such results mean that our approach can well locate the salient regions. Figure 9 shows the saliency maps across different frames in a same video, and we can see that our approach precisely identifies the saliency change due to mouth movement, while other approaches, especially [42], have almost no reaction to this type of movement. This confirms the effectiveness of our PF-DGMM model, which enables the dynamic transition of GMM between frames for modeling saliency of videos. Figure 10 further shows the saliency maps of different videos with faces of various sizes. Our approach is capable of predicting human attention well, regardless of face size. This further verifies the effectiveness of our approach in saliency detection of face videos.

Time complexity evaluation. The above performance evaluation has demonstrated that our approach performs better than other approaches in saliency detection accuracy of face videos. It is interesting to further compare the time complexity of our and other approaches. In our experiments, the computational time of saliency detection is recorded by running our and other approaches for 720p videos in Matlab 2014a and a computer with an I7-4771 @3.50GHz and 16G memory. Then, the computational time per frame is obtained and shown in Figure 11, for our and other approaches. We can see that our approach consumes more time than the early works of Cerf *et al.*

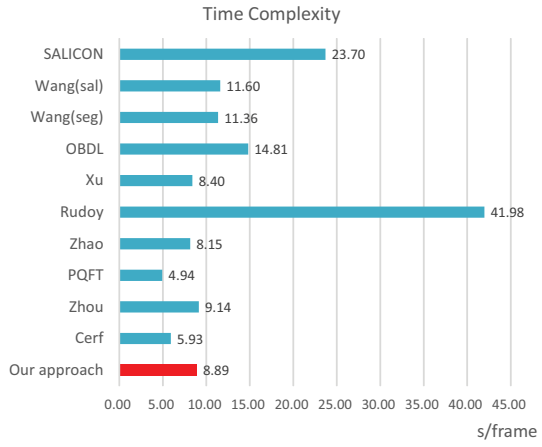


Fig. 11: Running time per frame of 720p, for our approach and other state-of-the-art approaches.

[39] and PQFT [21]. However, our approach performs much better than these approaches in saliency detection accuracy, as can be seen in Table II. More importantly, the computational time of our approach is comparable to or less than those of other state-of-the-art approaches, despite performing better than these approaches. This indicates efficacy and efficiency of our approach in saliency detection of face videos.

C. Evaluation on other databases

To test the generalizability of our approach, this section evaluate face videos from other databases. We selected totally 28 videos including one obvious face from the existing eye

TABLE II: The comparison of our, PF-DGMM channel, and other approaches in mean (standard deviation) of AUC, NSS, and CC, for all 46 test videos of our database.

Metrics	Our	Cerf [39]	Zhou [60]	PQFT [21]	Zhao [40]	Rudoy [34]	Xu [42]	OBDL [28]	WangSeg [14]	WangSal [11]	SALICON [32]
NSS	**5.20**(1.36)	2.41(0.59)	2.07(0.75)	1.16(0.80)	3.91(1.01)	2.15(0.82)	4.79(1.06)	1.58(1.21)	2.00(0.85)	1.73(1.00)	4.51(1.21)
CC	**0.84**(0.10)	0.58(0.10)	0.50(0.15)	0.25(0.15)	0.77(0.11)	0.57(0.13)	0.77(0.12)	0.31(0.14)	0.48(0.19)	0.42(0.22)	0.57(0.20)
AUC	**0.94**(0.03)	0.92(0.04)	0.88(0.06)	0.81(0.07)	**0.94**(0.03)	0.88(0.06)	0.93(0.06)	0.82(0.07)	0.58(0.21)	0.57(0.20)	0.76(0.12)

TABLE III: The comparison of our and other approaches in AUC, NSS, and CC, averaged over face videos of other database.

Metrics	Our approach	Cerf [39]	Zhou [60]	PQFT [21]	Zhao [40]	Rudoy [34]	Xu [42]	OBDL [28]	WangSeg [14]	WangSal [11]	SALICON [32]
NSS	**3.53**	1.91	1.30	1.00	2.63	2.36	3.19	1.31	1.45	1.63	2.73
CC	**0.64**	0.46	0.32	0.21	0.58	0.59	0.62	0.34	0.37	0.43	0.48
AUC	**0.91**	0.87	0.78	0.80	0.89	0.89	0.87	0.82	0.85	0.86	0.88

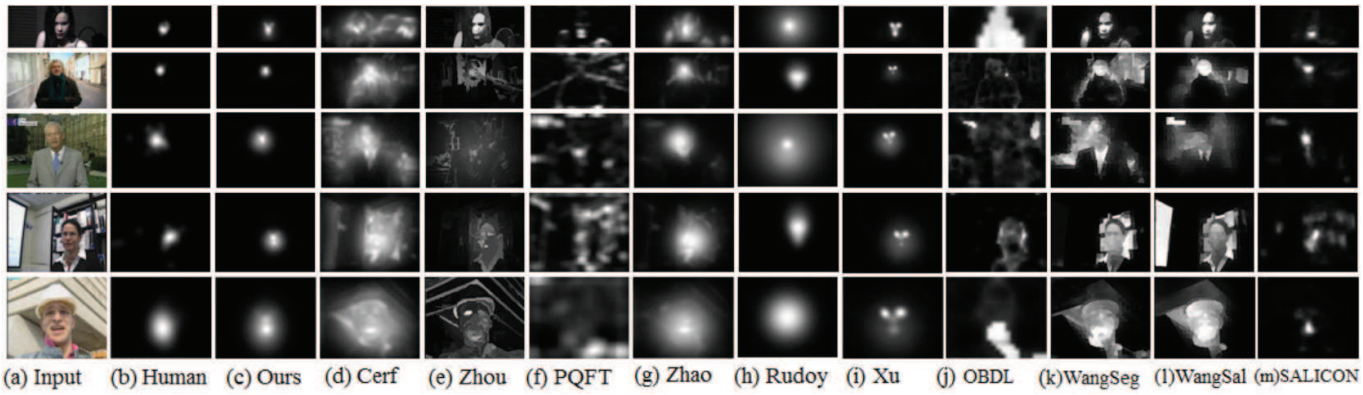


Fig. 12: Saliency maps of face videos selected from other databases, SFU [67], DIEM [68] and Hollywood [58].

tracking database of videos, SFU [67], DIEM [68] and Hollywood [58]. They are all tested in our experiment. Note that we use the same training set as Section VI-B, which includes 30 face videos.

Figure 12 shows the saliency maps of our and other 10 approaches, for the selected frames of four test videos. It is obvious that saliency maps of our approach are closer to ground-truth human fixation maps than other approaches. Table III further tabulates the NSS, CC and AUC results of our and other approaches, averaged over all 28 test videos of other databases. As aforementioned, it has been proved by [66] that AUC is not a robust metric for evaluating saliency, whereas NSS is the most robust one. From Table III we can find that our approach again performs better than all other approaches. Specifically, there is 0.34 NSS enhancement of our approach over the second ranking approach [42], similar to the 0.41 improvement in NSS on our database. Hence, the generalizability of our approach confirmed.

D. Performance analysis of our approach

As discussed in Section IV, PF-DGMM is a new feature channel proposed in our approach for modeling dynamic attention on face. Hence, it is interesting to analyze the performance of the single PF-DGMM channel for saliency detection in face videos. To this end, we evaluate the component quantitative results for our approach. Specifically, we evaluate NSS, CC and AUC of only applying the PF-DGMM channel in detecting saliency, averaged over all test videos in our database. For comparison, we also evaluate the performance of our approach without the PF-DGMM channel. The results are reported in Figure 13. We can see from this figure that the proposed PF-DGMM channel achieves 4.43 in NSS, 0.84 in CC and 0.92 in AUC, far better than those of our approach without the PF-DGMM channel. This indicates the effectiveness of the proposed PF-DGMM algorithm in dynamically modeling attention on face.

Figure 13 further shows the performance of each individual channel of color, intensity and orientation, which are also incorporated in our approach. We can see from this figure that the proposed PF-DGMM channel achieves considerably better performance than each individual channel, i.e., color, intensity and orientation. Additionally, Figure 13 shows the performance of our approach, which integrates the proposed PF-DGMM channel with other channels of color, intensity and orientation. We can see from this figure that when integrated with the channels of color, intensity and orientation, the performance of PF-DGMM can be improved from 4.43 to 5.20 for NSS, 0.78 to 0.84 for CC, and 0.92 to 0.95 for AUC. Thus, the effectiveness of feature integration can be validated.

Now, we analyze the failure cases of our approach in detecting saliency of face videos. Figure 14 shows four examples of failure cases, belonging to two videos. As seen in the second and third rows, the faces are missed by the face alignment method utilized in our approach. In this case, our approach can only apply the traditional feature channels of color, intensity and orientation, in saliency detection. However, face still attracts most visual attention, such that the saliency maps of our approach are far from the ground-truth attention maps. Therefore, more robust face alignment is required for succeeding in saliency detection of face videos. We can further see from the first row of Figure 14 that the lady wearing sunglass attracts little attention in the eye regions, due to the

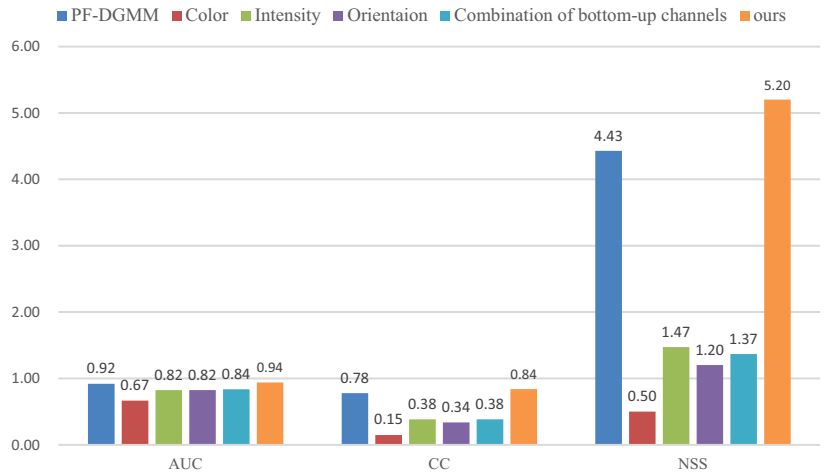


Fig. 13: AUC, CC, and NSS results for single channels (PF-DGMM, color, intensity and orientation) and the combination of three bottom-up channels (color, intensity and orientation).

occlusion of eye regions. Although face alignment applied in our approach is able to detect the face region, our saliency detection approach still fails in this case. It is because the eye regions are with some kind of saliency in our approach, not in accordance with ground-truth attention map. The last row of Figure 14 shows that the occlusion of mouth by the hand makes fixation distribution not follow Gaussian model. In this case, our DGMM distribution of saliency does meet the distribution of ground-truth attention on the occluded mouth region. For this case, it is necessary to study the influence of face occlusion on the distribution of visual attention.

VI. Conclusions

In this paper, we have proposed a promising data-driven saliency detection approach for face videos, which can generate accurate face saliency by taking into account the face size and mouth movement. Specifically, we set up a new eye tracking database, which is composed of 76 videos viewed by 40 subjects. Then, four observations were made from our database, implying that the DGMM is suitable for modelling the distribution of visual attention on face videos. Inspired by these observations, DGMM was developed to predict saliency of face videos by learning from eye tracking data. In addition, the PF algorithm was employed in our approach to sequentially update the parameters of DGMM along with processed video frames. Thus, our approach is called PF-DGMM, which can seen as a data-driven approach. At last, the experimental results demonstrated that our approach can more accurately detect the visual saliency of face videos, compared with other state-of-the-art approaches.

Our work, at the current stage, mainly focuses on detecting saliency of the face without any occlusion. When the face is occluded in videos, our approach fails in modelling attention as the occlusion alters the distribution of human attention. Thus, it is rather an interesting future work to study the influence of face occlusion on visual attention. The proposed approach may facilitate the future research in the area of the video analysis on emotional behavior, by considering the more attractive regions in face. Upon the study of this paper, it is expected that the coding efficiency of face videos can be further improved by removing the perceptual redundancy existing in the non-salient regions. In this way, we can use fewer bits to encode and transmit face videos, to relieve the bandwidth-hungry dilemma caused by the prevalence of video conferencing applications, e.g., FaceTime, Skype, and even Wechat.

References

[1] 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.

[2] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 10, pp. 3137–48, 2015.

[3] L. Chen, J. Shen, W. Wang, and B. Ni, "Video object segmentation via dense trajectories," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2225–2234, 2015.

[4] W. Wang, J. Shen, and F. Porikli, "Selective video object cutout," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[5] W. Wang, J. Shen, H. Sun, and L. Shao, "Vicos2: Video co-saliency guided co-segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, 2017.

[6] L. Zhang, M. Wang, L. Nie, R. Hong, Y. Xia, and R. Zimmermann, "Biologically inspired media quality modeling," in *ACM international conference on Multimedia (ACM MM)*, 2015, pp. 491–500.

[7] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *IEEE Transactions on Multimedia*, vol. 8, no. 1, pp. 11–18, 2006.

[8] W. Wang, J. Shen, Y. Yu, and K. L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[9] L. Huo, L. Jiao, S. Wang, and S. Yang, "Object-level saliency detection with color attributes," *Pattern Recognition*, vol. 49, pp. 162–173, 2016.

[10] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5025–5034, 2016.

[11] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 11, pp. 4185–4196, 2015.

[12] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 26, no. 7, p. 3156, 2017.

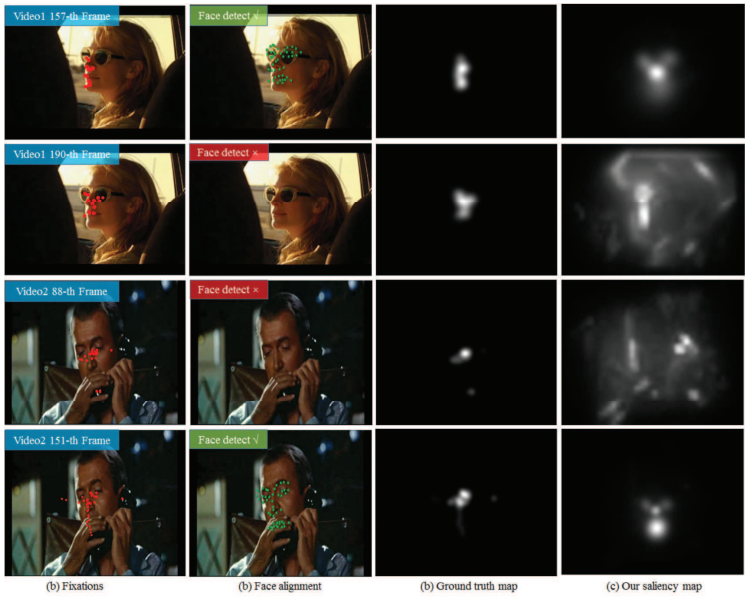


Fig. 14: Failure cases of our approach. The red dots of the first column images are human fixations. The green dots in the second column are the facial landmarks.

[13] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 8, pp. 2303–2316, 2015.

[14] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[15] 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.

[16] 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.

[17] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Optical Science and Technology, SPIE's 48th Annual Meeting*. International Society for Optics and Photonics, 2004, pp. 64–78.

[18] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[19] G. Boccignone, "Nonparametric bayesian attentive video analysis," in *ICPR*. IEEE Computer Society Press, 2008, pp. 1–4.

[20] L. Zhang, M. H. Tong, and G. W. Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proceedings of the 31st Annual Cognitive Science Conference*. AAAI Press Cambridge, MA, 2009, pp. 2944–2949.

[21] 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*, vol. 19, no. 1, pp. 185–198, 2010.

[22] 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.

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

[24] 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.

[25] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 2, pp. 314–328, 2013.

[26] W. Hou, X. Gao, D. Tao, and X. Li, "Visual saliency detection using information divergence," *Pattern Recognition*, vol. 46, no. 10, pp. 2658–2669, 2013.

[27] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 1, pp. 27–38, 2014.

[28] 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.

[29] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Human visual system-based saliency detection for high dynamic range content," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 549–562, 2016.

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

[31] 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.

[32] 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.

[33] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *CVPR*, June 2016.

[34] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *CVPR*, 2013, pp. 1147–1154.

[35] C. Chen, S. Li, H. Qin, and A. Hao, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognition*, vol. 52, pp. 410–432, 2016.

[36] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150–165, 2010.

[37] U. Rajashekar, I. Van Der Linde, A. C. Bovik, and L. K. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *Image Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 564–573, 2008.

[38] 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.

[39] 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.

[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] M. Xu, Y. Ren, and Z. Wang, "Learning to predict saliency on face images," in *ICCV*, 2015, pp. 3907–3915.
[43] C. M. Bishop, *Pattern recognition and machine learning*. springer New York, 2006.
[44] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of h. 264/avc," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 1, pp. 134–139, 2008.
[45] M. Xu, Y. Liu, Z. Wang, and H. Hu, "Find who to look at: Turning from action to saliency," in *Submitted to IEEE transactions on image processing*, 2016.
[46] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house Boston, 2004, vol. 685.
[47] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.
[48] M. Kmmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *Computer Science*, 2014.
[49] S. Jetley, N. Murray, E. Vig, undefined, undefined, undefined, and undefined, "End-to-end saliency mapping via probability distribution prediction," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, pp. 5753–5761, 2016.
[50] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.
[51] N. D. B. Bruce, C. Catton, and S. Janjic, "A deeper look at saliency: Feature contrast, semantics, and beyond," in *Computer Vision and Pattern Recognition*, 2016, pp. 516–524.
[52] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Computer Vision and Pattern Recognition*, 2016, pp. 5781–5790.
[53] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.
[54] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *ICCV*, vol. 1, 2001, pp. I–511.
[55] J. Zhao, C. Siagian, and L. Itti, "Fixation bank: Learning to reweight fixation candidates," in *Computer Vision and Pattern Recognition*, 2015, pp. 3174–3182.
[56] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 893–907, 2017.
[57] Y. Kavak, E. Erdem, and A. Erdem, "A comparative study for feature integration strategies in dynamic saliency estimation," *Signal Processing: Image Communication*, vol. 51, pp. 13–25, 2017.
[58] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conference on Computer Vision*, 2012, pp. 842–856.
[59] T. Mauthner, H. Possegger, G. Waltner, and H. Bischof, "Encoding based saliency detection for videos and images," in *Computer Vision and Pattern Recognition*, 2015, pp. 2494–2502.
[60] F. Zhou, S. B. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3358–3365.
[61] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 50–58.
[62] J. Saragihand, S. S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *ICCV*, 2009, pp. 1034–1041.
[63] J. Anderson and B. Anderson, "The myth of persistence of vision revisited," *Journal of Film and Video*, vol. 45(1), pp. 3–12, 1993.
[64] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[65] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, vol. 12, no. 656-704, p. 3, 2009.
[66] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen, "A data-driven metric for comprehensive evaluation of saliency models," in *ICCV*, 2015.
[67] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," *Image Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 898–903, 2012.
[68] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, Mar. 2011.

Mai Xu (M'10, SM'16) received B.S. degree from Beihang University in 2003, M.S. degree from Tsinghua University in 2006 and Ph.D degree from Imperial College London in 2010. From 2010-2012, he was working as a research fellow at Electrical Engineering Department, Tsinghua University. Since Jan. 2013, he has been with Beihang University as an Associate Professor. During 2014 to 2015, he was a visiting researcher of MSRA. His research interests mainly include image processing and computer vision. He has published more than 60 technical papers in international journals and conference proceedings, e.g., IEEE TIP, CVPR and ICCV. He is the recipient of best paper awards of two IEEE conferences.

Yun Ren (S'15) received the bachelors degree in Beihang University, Beijing, China, in 2015, where she is currently pursuing the Master degree. She is admitted and expected to receive the Master degree in Mar. 2018. During her study, she was also awarded 10 scholarships, including National Scholarship, and nearly 20 competition prizes. Her research interests include saliency prediction and video analysis.

Zulin Wang (M'14) received the B.S. and M.S. degrees in electronic engineering from Beihang University, in 1986 and 1989, respectively. He received his Ph.D. degree at the same university in 2000. He is currently the dean of school of electronic and information engineering, at Beihang University, Beijing, China. His research interests include image processing, electromagnetic countermeasure, and satellite communication technology. He is author or co-author of over 100 papers and holds 6 patents, as well as published 2 books in these fields. He has undertaken approximately 30 projects related to image/video coding, wireless communication, etc.

XiaoMing Tao (M'09) received the B.E. degree from the school of Telecommunications Engineering, Xidian University in 2003, and Ph. D. from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2008. She is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. Her research interests include wireless communications and networking, and multimedia signal processing.