Find Who to Look at: Turning From Action to Saliency

Mr. G Ahmed Zeeshan[1] , Mrs. Nuzath Unnisa Begum[2] , Mr. Sanka Ravi[3]

1,2 Associate Professor,3 Assistant Professor,
1,2,3 Department of ECE,
1,2,3 Global Institute of Engineering & Technology, Moinabad, Rangareddy Dist., Telangana State.

Abstract—

The past decade has witnessed the use of highlevel features in saliency prediction for both videos and images. Unfortunately, the existing saliency prediction methods only handle high-level static features, such as face. In fact, high-level dynamic features (also called actions), such as speaking or head turning, are also extremely attractive to visual attention in videos. Thus, in this paper, we propose a data-driven method for learning to predict the saliency of multiple-face videos, by leveraging both static and dynamic features at high-level. Specifically, we introduce an eye-tracking database, collecting the fixations of 39 subjects viewing 65 multiple-face videos. Through analysis on our database, we find a set of high-level features that cause a face to receive extensive visual attention. These high-level features include the static features of face size, center-bias and head pose, as well as the dynamic features of speaking and head turning. Then, we present the techniques for extracting these high-level features. Afterwards, a novel model, namely multiple hidden Markov model (M-HMM), is developed in our method to enable the transition of saliency among faces. In our M-HMM, the saliency transition takes into account both the state of saliency at previous frames and the observed high-level features at the current frame. The experimental results show that the proposed method is superior to other state-of-the-art methods in predicting visual attention on multiple-face videos. Finally, we shed light on a promising implementation of our saliency prediction method in locating the region-of-interest, for video conference compression with high efficiency video coding.

Index Terms— Video analysis, saliency prediction, face.

Manuscript received June 9, 2017; revised May 10, 2018; accepted May 11, 2018. Date of publication May 16, 2018; date of current version June 15, 2018. This work was supported in part by the Natural Key R&D Program of China under Grant 2017YFB1002400, in part by NSFC projects under Grant 61573037, in part by the Fok Ying-Tong Education Foundation under Grant 151061, in part by the Zhejiang Public Welfare Research Program under Grant 2016C31062, and in part by the Natural Science Foundation of Zhejiang Province under Grant LY16F010004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Scott T. Acton. (Corresponding author: Feng He.)

M. Xu and F. He are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail:maixu@buaa.edu.cn; robinleo@buaa.edu.cn).

Y. Liu is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China, and also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: fannie_lyf@buaa.edu.cn).

R. Hu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: haoji_hu@zju.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a PDF with the analysis of inter-subject consistency in attractiveness rating. The total size of the file is 0.0199 MB. Contact maixu@buaa.edu.cn for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org. Digital Object Identifier 10.1109/TIP.2018.2837106

I. INTRODUCTION

A. Background WHEN people are exposed to a large scene, they use their fovea to perceive an area of interest with high resolution. The other regions, namely the peripheral regions, are perceived with low resolution. Therefore, under the limitation of humans brain processing capacity, visual attention enables humans to effectively process considerable amounts of visual data [1]. Over the past decades, visual attention modeling has been broadly studied in the fields of neurophysiology, computer vision and multimedia [2]. Saliency prediction is an effective way to model the deployment of possible visual attention on images or videos. Recently, saliency prediction has been widely applied in object detection [3], image retargeting [4], visual quality assessment [5] and video coding [6]. B. Related Work Saliency prediction can be traced back to Itti's model [7], which combines the center-surround features of color, intensity and orientation together. However, Itti's model [7] mainly focuses on images. For video saliency prediction, the initial work is [8], in which Itti's model was extended by incorporating two dynamic features, i.e., motion and flicker contrast. Both [7] and [8] are low-level based methods, which explore and integrate some low-level features for saliency detection. Afterwards, low-level based video saliency prediction evolves alongside directions of feature exploration and feature integration. In exploring saliency-related features, surprise is defined in [9] as the Kullback-Leibler divergence (KL) between spatiotemporal posterior and prior beliefs across video frames. Then, a Bayesian framework was developed in [9] to calculate surprise for predicting video saliency. Besides, sparse representation of learnt texture atoms (SR-LTA) was proposed in [10] as low-level features to predict saliency, benefiting from the recent success of dictionary learning. Besides, some compressed domain features, such as motion vector in [11] and bit allocation in [12], were also utilized as low-level features for low-level based video saliency prediction. In integrating saliency-related features, some advanced works were proposed. In particular, a graph-based visual saliency (GBVS) was proposed in [13] for saliency prediction, which applies graph model in combing low-level features of color, intensity and orientation. There also exist dynamic saliency models [14] and [15] fusing spatio and temporal visual features to generate saliency maps. Later, Guo and Zhang [16] proposed to integrate four low-level features (two color features, one intensity

feature and one motion feature) using the phase spectrum of quaternion Fourier transform (PQFT). Most recently, support vector machine (SVM) [17] has been utilized for learning to integrate low-level features in video saliency prediction. However, the relationship between low-level features and human visual attention is rather complicated, as the understanding of the HVS is still in its infancy. On the contrary, high-level features (e.g., object, text and face) are the more evident cues to receive a great amount of visual attention. Thus, a large number of methods have recently employed highlevel features for the saliency prediction of images [18]–[24], and these methods can be seen as high-level based methods. Those high-level based methods can be classified into the saliency prediction of generic images and face images. For generic saliency prediction, Judd et al. [18] combined high-level features (e.g., face and text), middle-level features (e.g., gist) and low-level features together, via learning their corresponding weights with SVM. Most recently, Huang et al. [19] have proposed the saliency in context (SALICON) method to incorporate the high-level semantic features of objects in

saliency prediction, in light of deep neural networks (DNN). Similarly, Bruce et al. [25] proposed a fully convolutional networks (FCN) based model to automatically extract high-level features in saliency prediction and salient object segmentation. In addition, Shao et al. [26] used DNN to extract semantic features fusing with low-level features and saccadic amplitude to predict scanpath. For face images, Cerf et al. [20] proposed to add face as an additional feature into Itti's model [7], such that the saliency prediction accuracy can be dramatically improved. The impact of face in the saliency prediction of face images was further investigated in [21]. Later, Xu et al. [22] proposed to precisely model saliency of face region, via learning the fixation distributions of face and facial features. Meanwhile, Jiang et al. [23] developed several face-related features at high-level to predict saliency in a scene with multiple faces. These high-level features include face size, pose and location. There have also emerged some high-level based methods [27]–[30] that make use of high-level features, for video saliency prediction. Specifically, Pang et al. [27] proposed to explore the high-level based information of eye movement patterns, i.e., passive and active states [31], to model attention on videos. Later, Hua et al. [28] proposed to learn middlelevel features, i.e., gists of a scene, as the high-level based cues in video saliency prediction. Rudoy et al. [29] proposed to predict the saliency of a given frame, conditioned on the detected saliency of previous reference frames. In their method, high-level features (e.g., people) and low-level features are integrated to perform saliency prediction for currently processed frames. In [30], the high-level feature of camera motion was incorporated for video saliency prediction. Most recently, DNN has been developed in [32] and [33] for learning some high-level features to predict video saliency. The saliency prediction of face images has been extensively studied in [20]–[23]. Similarly, several works [34]–[38] have been devoted to saliency prediction of face videos, which focus on talking face and consider the influence of sound on visual attention. However, most of them only concentrate on the conversation videos and do not aim at predicting the salient face among multiple faces.

Fig. 1. Examples of visual attention (viewed by 39 subjects) on multiple-face videos influenced by different actions. Each row shows one video with their attention heat maps. Some selected frames of these videos are provided in each column. In the first and second columns, visual attention is attracted by the action of head turning (profile-to-front and front-to-profile). In the third column, the action of speaking receives substantial visual attention. Note that the videos are chosen from our database, to be discussed in Section II.

In fact, it is intuitive that some high-level dynamic features, also called actions, may attract extensive visual attention in a face video. For example, Figure 1 illustrates that most attention is focused on one face, related to the actions of speaking or head turning. Unfortunately, to our best knowledge, few existing video saliency prediction methods consider the impact of multiple high-level dynamic features on visual attention, despite single high-level dynamic feature of speaking being well embedded in those methods [34]–[38]. It is worth mentioning that most recently, human actions have been explored [39] to find the key person for event detection in videos of basketball games, in the area of recognition. However, the prediction of the key person does not produce saliency, because the

correlation between the detected key person and ground truth attention is not investigated . Moreover, it is limited to basketball videos with human bodies. C. Our Work and Main Contributions In this paper, we propose a novel method to predict the saliency of multiple-face videos, by modeling temporal transition of saliency with regard to high-level static and dynamic features. We found out that the most popular videos of YouTube contain dialogue scenes (such as TV programs, movies, etc), including one or more faces. Thus, this paper mainly concentrates on multiple-face videos, in which faces and their high-level dynamic features are indeed useful in determining saliency as illustrated in Figure 1. It is worth pointing out that the demand on video conferencing, like FaceTime and Skype, is undergoing the growth explosion, posing the bandwidth-hungry issue. To relieve this issue, this paper discusses a potential implementation of our method in high efficiency video coding (HEVC) [40] of video conferencing, which can improve subjective quality at limited bit-rates via locating a salient face as the region-of-interest (ROI). Specifically, we established an eye-tracking database, which is comprised by fixations of 39 subjects viewing 65 multipleface videos. We mine our database to investigate how important the high-level static/dynamic features are in drawing

TABLE I
VIDEO CATEGORIES IN OUR DATABASE

Category	TV play/movie	group interview	individual interview	video conference	variety show	music/talk show	group discussion	overall
Number of videos	12	12	8	6	7	10	10	65

Fig. 2. One example for each category of videos. From left to right, the videos belong to TV play/movie, group interview, individual interview, video conference, variety show, music/talk show, and group discussion.

visual attention. Our investigation revealed that most of human attention is attracted by one among multiple faces in a video, which is correlated with the size, center-bias and pose of the face (seen as high-level static features). These features are thus leveraged in our method as the high-level static features for predicting the visual attention of each video frame. This is similar to the work of [23], which refers to saliency prediction among multiple faces in images. Beyond [23], we find that the high-level dynamic features of speaking and head turning attract even more visual attention, and hence, they are utilized as high-level dynamic features for videos. Then, we propose a multiple hidden Markov model (M-HMM) to predict the dynamic transitions of saliency between faces across video frames, according to the above high-level features (either static or dynamic). The difference between [23] and our method is that [23] is proposed for predicting the saliency of multiple-face images with only high-level static features, whereas our method aims at applying M-HMM to predict the saliency transition of multiple-face videos upon both static and dynamic features. In summary, we make four contributions in this paper. (1) We argue that high-level static and dynamic features can draw extensive attention in multiple-face videos, based on a thorough analysis using our eye-tracking database. (2) We develop techniques to extract the actions of speaking and head turning, as the high-level dynamic features for saliency prediction. (3) We propose an M-HMM method to take advantage of observed high-level features, achieving the temporal transition of saliency across multiple faces in videos. (4) We provide a promising implementation of our saliency prediction method, locating a salient face as the ROI for video conferencing coding. II. DATABASE ESTABLISHMENT This section describes how we conducted the eye-tracking experiment to establish our database, which is comprised by fixations of 39 subjects viewing 65 multiple-face videos. Our eye-tracking database is specialized for multiple-face videos. First, we asked 3 volunteers to randomly find videos from YouTube and Youku, with the criterion that the videos should contain obvious faces. Then, a set of 65 videos at 720p were collected, which contain various numbers of faces varying from 1 to 27. All of

these videos were compressed using H.264. The duration of each video was cut down to be around 20 seconds. Note that these 65 videos are with either indoor or outdoor scenes, and they can be classified into 7 categories1 (see Table I and Figure 2 for more details). Also note that the audio track is removed in our database and eye-tracking experiment, to make our approach focus on visual cues of saliency. Next, 39 subjects (26 males and 13 females, aging from 20 to 49), with either corrected or uncorrected normal eyesight, participated in our eye-tracking experiment to watch all 65 videos. Among these subjects, two were experts working in the field of saliency prediction. The other subjects did not have any experience on saliency prediction, and they were also naive to the purpose of our eye-tracking experiment. The eye fixations of the 39 subjects on viewing each video were recorded by a Tobii X2-60 eye tracker at 60 Hz. For the eye tracker, a 23-inch LCD screen was used to display the test videos at their original resolutions. During the eye-tracking experiment, all subjects were required to sit on a comfortable chair with the viewing distance being ∼60 cm from the LCD screen. Before viewing videos, each subject was required to perform a 9-point calibration for the eye tracker. Subsequently, the subjects were asked to free-view videos displayed at random order. In order to avoid eye fatigue, the 65 test videos were divided into 3 sessions, and there was a 5-minute rest after viewing each session. Moreover, a 10-second blank period with black screen was inserted between two successive videos for a short rest. Finally, the eye-tracking data on viewing all 65 videos were collected for our database, containing 1,011,647 fixations in total. For facilitating future research, our database is available online: https://github.com/yufanLiu/find. III. DATA ANALYSIS In Section I, we have shown the intuition that face, together with its high-level features, is an evident cue to attract visual attention in a multiple-face video. In this section, we thoroughly analyze the collected eye-tracking data of our database, to further predict the visual attention on multipleface videos. According to the analysis, several observations are investigated, to be discussed in the following. Note that the landmarks, features and actions of faces (i.e., speaking and head turning) for the following observations are manually annotated.2 The annotation results of all videos in our database are also downloadable, together with our eye-tracking results.

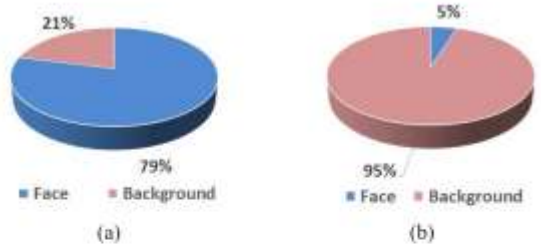
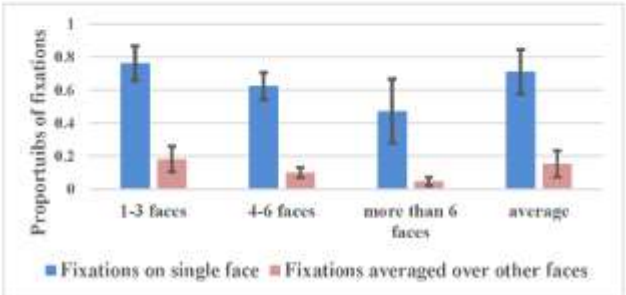


Fig. 3. Proportions of fixations and pixels in face and background over all 65 videos of our database.



A.Face vs. Attention Observation 1: In multiple-face videos, faces draw a significant amount of attention. At each video frame, the attention of different subjects consistently focuses on one face among all faces. Figure 3 shows the proportions of fixations and pixels belonging to face and background, in our database. We can see from this figure that despite taking up only 5% of the pixels, faces receive 79% of the fixations. This verifies that faces attract almost all visual attention in multiple-face videos. Figure 4 further plots the proportions of fixations falling into one face and the sum of those falling into other faces. We can conclude from this figure that human attention of different subjects is consistent in being attracted by one face among all faces. Besides, the subjective examples presented in Figure 1 also imply that faces, normally one face, draw most visual attention in a video. Meanwhile, there are only 14% of the fixations falling into torso and limbs. This implies that face attracts considerably more attention than the regions of torso and limbs. Observation 2: The amount of attention on each face has a small positive correlation with face size. Does the largest face receive more fixations than other faces in a video frame? To answer this question, we measure the correlation between the ranking of face size3 in a video and the corresponding saliency, via Spearman rank correlation coefficients [41]. Note that the Spearman correlation coefficient is a nonparametric measure of rank correlation. We also report the Pearson correlation coefficient results in 3Here, the size of each face is calculated by the number of pixels of the face region. In this paper, the face region is determined by contours of facial landmarks. the following analysis, to further verify our observations. The Spearman rank correlation coefficients and Pearson correlation coefficients are calculated according to the fixation number and face size of each face in a video frame. Then, the Spearman rank correlation coefficient and Pearson correlation coefficient of all frames, averaged over the 65 videos in our database, are 0.25 (p-value $p = 0.039$) and 0.32 ($p = 0.016$), respectively. Therefore, the positive correlation values suggest that a larger face may draw more attention, which is consistent with [42]. B. Static Features vs. Attention Observation 3: Humans are more likely to fixate on the face that is close to the video center, among all the faces at a video frame. The center-bias [2], [43] is an obvious cue to predict human fixations on generic videos. It is also intuitive that people are likely to pay their attention on the face that is close to the video center. We hence investigate the correlation of attention on a face with the Euclidean distance of this face to the video center. To quantify such correlation, we evaluate the average Spearman rank correlation coefficient ($\rho = -0.22$, $p = 0.019$) and Pearson correlation coefficient ($\gamma = -0.19$, $p = 0.007$), following the same way as Observation 2. The negative values of ρ and γ indicate that humans probably fixate on the face that is close to the video center. According to [44], human attention on the center face is mainly due to the photographer bias, which means that the photographer or video editor normally places the important face near the center of the video. Observation 4: In multiple-face videos, visual attention on each face is correlated with its head pose. One observation to explore is the relationship between visual attention and head pose for each face in multiple-face videos. In this paper, we define head pose by two categories: front and profile. Front is one case of pose that the angle between face-viewing and image plane is less than 25° . Profile is the other case of pose that the angle is in the range of [25°, 90°]. There are in total 110,544 frontal faces and 30,007 profile faces in our database. Figure 5 shows that the frontal face is more attention-capturing than the profile face in a video frame. We further find that when speaking, frontal faces receive 12.6 fixations per face, while profile faces only draw 7.8 fixations per face. Observation 5: Visual attention is almost irrelevant to face attractiveness. One hypothesis is that the attention on different faces in a multiple-face video may be relevant to face aesthetic. We therefore analyze this relevance. We follow the way of [45] to measure the attractiveness of faces. Twenty-eight subjects participated in rating the attractiveness of each face, over all 65 videos in our database. The rating score ranges from 1 to 10, and a larger score means a more beautiful face. Then, the scores of all 28 subjects are averaged to obtain the attractiveness value of each face. We find that the average Spearman rank correlation coefficient is $\rho = 0.05$ with $p =$

0.266, as the correlation between attention and face attractiveness. The corresponding Pearson correlation coefficient is $\gamma = -0.03$ with $p = 0.268$. Surprisingly, visual attention is almost irrelevant to face attractiveness. This is

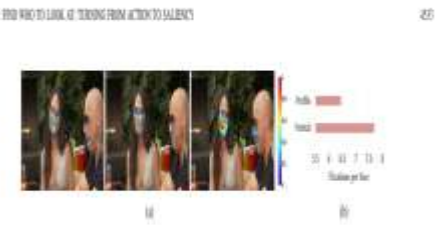


Fig. 5. Comparison of attention in front and profile faces. Note that (a) is the results of three frames of a randomly selected video. Also, note that the statistical results in (b) are averaged over the fixation data of all 65 videos in our database. In (b), fixations per face are shown for frontal and profile faces, respectively

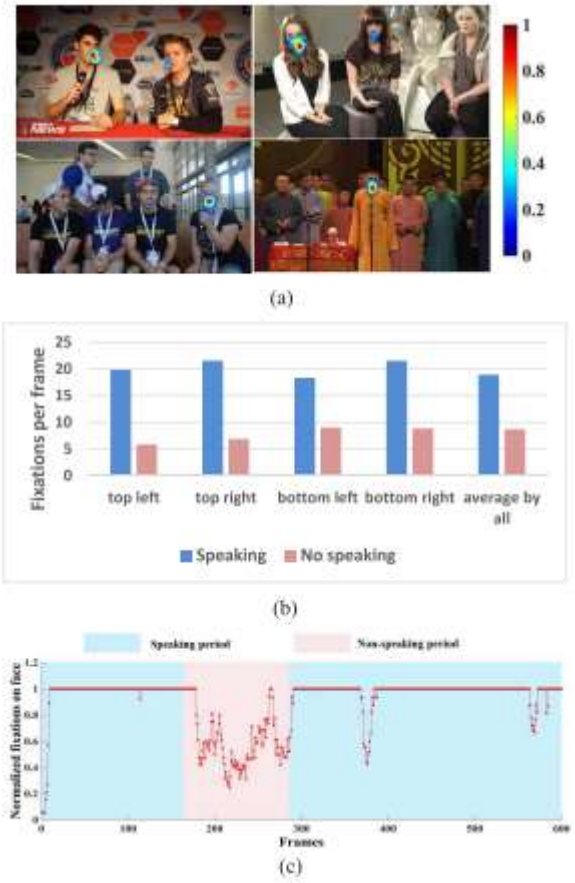


Fig. 6. Human fixations in speaking and non-speaking faces. (a) is the fixation maps of 4 randomly selected videos at different crowd levels, containing 2, 3, 6, and 10+ persons. (b) shows the numbers of fixations per frame in speaking and non-speaking faces, for each individual video of (a). In (b), the bar of "average by all" shows the numbers of fixations per face, averaged over all speaking and non-speaking faces of all 65 videos in our database. (c) shows the actions of speaking and non-speaking versus normalized fixations of one face of a selected video. In (c), fixations are normalized, by dividing the fixation number of each face with the maximal fixations among all faces. probably due to the fact that visual attention is normally drawn by face actions, as revealed in the following

observations. C. Dynamic Actions vs. Attention Observation 6: A speaking face attracts a large amount of visual attention. Figure 6 shows the relationship between the action of speaking and the fixations in multiple-face videos. We can see from the subjective results in Figure 6-(a) that human tends to look at the speaking face. Note that the interview-like videos (with microphones) are chosen as examples, because the microphones in these videos help readers locate the speaking face. Figure 6-(b) quantifies the numbers of average fixation on speaking and non-speaking faces, for the examples of Figure 6-(a). More importantly, the statistical results of "average by all" in Figure 6-(b) are averaged over all 65 videos in our database, which verifies that speaking action attracts approximately 20 fixations per frame, whereas non-speaking action attracts less than 9 fixations per frame. Figure 6-(c) also plots the actions of speaking and non-speaking versus visual attention for a video. In summary, we can observe from Figure 6 that the speaking action (i.e., mouth motion) may draw extensive visual attention to the corresponding face in multiple-face videos. Observation 7: In multiple-face videos, visual attention on each face is highly correlated with head turning. It is also interesting to find out the correlation between visual attention and head turning, in multiple-face videos. Figure 7-(a) illustrates that fixations drop when head turns from front to profile, and that attention increases when head turns from profile to front. Note that the statistical results of Figure 7-(a) are obtained by averaging over all videos in our database. Figure 7-(b) provides some examples to show how visual attention is attracted by head turning. We can observe from Figure 7 that the front-to-profile head turning significantly reduces visual attention, while the profile-to-front head turning receives increasing visual attention. IV. FEATURE DETECTION Since Section III has found that visual attention is highly correlated with some high-level features of face, this section mainly discusses the techniques for detecting these features. Specifically, Section IV-A describes the preliminary for facerelated feature detection, including tracking faces and their landmarks in videos. After tracking facial landmarks, the size and center-bias of face can be easily obtained. Section IV-B proposes a technique to monitor the action of speaking. Section IV-C presents a way to detect features of head pose and head turning. A. Preliminary Observation 1 verified that a face is an obvious cue to draw visual attention in a video. Accordingly, we need to detect faces in multiple-face videos. Additionally, the landmarks of faces are necessary to detect high-level features, such as speaking. Thus, this section concentrates on the detection of face and facial landmarks for multiple-face videos, as the

Fig. 7. Correlation between fixations and head turning. Fixation change per video averaged over all 65 videos in our database, when head turns from front to profile ($F \rightarrow P$) and from profile to front ($P \rightarrow F$). (b) Fixation maps for the frames of head turning.

preliminary of our saliency prediction method. The recent work of [46] constructed a unified model for face detection, pose estimation and landmark estimation, in multiple-face images. Here, we first utilize [46] to detect faces and their landmarks at each frame of a video, in which both frontal and profile faces can be located. To improve face detection performance, we follow our recent work [47] to manage some harsh situations, such as partial occlusion and poor light conditions, by exploring

temporal information of videos. To be more specific, we match the faces across frames, by searching for the face with nearest Euclidean distance. We then identify the nearest faces of two consecutive frames as the matched face of the same person, provided that their distance is less than a threshold:

$$th_E = \gamma \times \sqrt{w^2 + h^2}, \qquad (1)$$

where w and h are the width and height of the detected face, respectively. Otherwise, we regard them as non-matching faces, belonging to different persons. In (1), γ is a parameter to control the sensitivity of face matching, and it is simply set to 0.5 in this paper. When matching faces across frames, some faces may be missed due to occlusion or light conditions. For detecting these missed faces, the linear interpolation of faces is applied to neighboring frames within a sliding window. In this paper, the length of the sliding window is empirically chosen to be 17, to make the face detection results appropriate. The experimental results have verified that the above technique is simple yet effective in matching faces of our database, which can also handle camera motion; thus, it is not necessary to utilize another advanced tracking algorithm. Next, we also use [46] to locate facial landmarks in multiple-face videos. In our method, [46] is directly used to locate 39 landmarks for profile faces. Then, we improve the performance of [46] in landmark localization for frontal faces, via applying the latest work of [48] to track landmarks for each detected frontal face. After faces are interpolated in some video frames, we implement our previous work of [47] to predict the facial landmarks upon the matched faces of neighboring frames. As a result, multiple faces, either frontal or profile, can be detected and matched in a video with well-located landmarks. Finally, the size and center-bias of each face should be estimated using facial landmarks in a video, since Observations 2 and 3 have shown that attention is correlated with the size and center-bias of face. Specifically, the contour and region of each face are extracted by connecting the related landmarks. Then, the number of pixels belonging to the face region is considered to be the face size. Based on the contour of the extracted face, the face center can also be estimated, and its Euclidean distance to the video center is calculated as the center-bias of each face. Note that both the size and center-bias of each detected face should be normalized by video resolution. In addition, the performance of our saliency prediction method relies on the results of the above face detection and tracking algorithm, which is the basis of our method. B. Detection on Speaking and Non-Speaking Observation 6 has shown that speaking may attract a large amount of visual attention. Thus, we now present an algorithm to detect the actions of speaking. The procedure of our algorithm is summarized in Figure 8, and it learns to detect the speaking action using the motion, geometry and texture of mouth regions. In general, we first incorporate a classic motion detection approach, optical flow [49], to measure the intensity and orientation of mouth motion. Second, we leverage the detected mouth landmarks to measure the elongation of the mouth for quantifying the geometry variation of speaking. Third, the gray scale value of the mouth region pixels is utilized to find the texture variation of speaking, similar to [50] and [51]. Finally, our algorithm applies SVM as the binary classifier of speaking, with respect to the features of optical flow, mouth elongation and gray values. Specifically, the geometry of the mouth variation is used as a feature to make a judgement on speaking. Toward such a geometry, the height and width of outer and inner lips are measured on the basis of mouth landmarks. We define the height and width of the outer lip by a and b, respectively, and the height and width of the inner lip are denoted as c and d, respectively. Refer to Figure 9 for more details. Then, the elongation of the mouth can be calculated by V = a + c b + d . (2) Also, the texture change of the mouth region is incorporated in speaking detection. The previous work of [51] has found that speaking may change the distribution of gray values in the mouth region. Specifically, if most pixels of mouth region

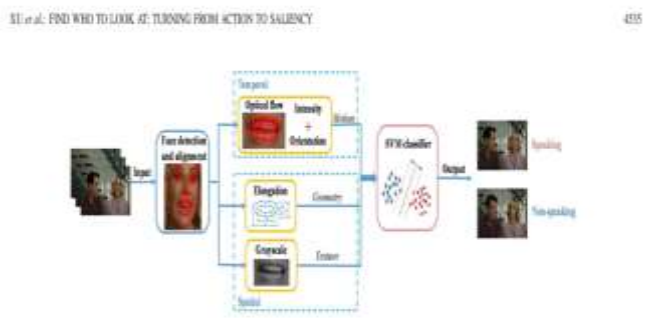


Fig. 8. Framework of the speaking detection algorithm. In this framework, [46] is applied for face detection and alignment, such that both frontal and profile faces can be processed. Likewise, there are 68 and 39 landmarks for frontal and profile faces, respectively. For profile faces, the calculation of elongation is different, which uses different landmarks to compute the corresponding variables.

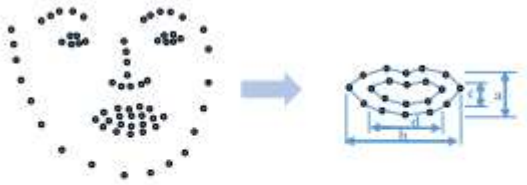


Fig. 9. Illumination for the height and width of outer and inner lips by facial landmarks. Left is the facial landmark graph, and right is the landmarks of the mouth.

are at very low gray scale, the person is more likely to speak. It is because when speaking, mouth cavity decreases the average intensity of mouth region due to black region. Here, we follow [51] to use the gray values of the mouth region as one feature for speaking detection. The binary process is conducted on the gray image of the mouth region, with regard to a predefined threshold thG . Then, the average binary value of the mouth region is computed by

$$B = \frac{\sum_{(x,y)\in \mathbf{R}} b(x,y)}{\#(\mathbf{R})}, \qquad (3)$$

where #(R) is the total number of pixels in the mouth region R, and b(·) is the binary value of each pixel in the mouth region. Next, we estimate the intensity of mouth motion based on optical flow. Here, the mouth region in a video frame, defined by R, is extracted by connecting landmarks of the outer lips. In the mouth region, we apply the Lucas-Kanade algorithm [49] to detect pixel-wise optical flow. Then, the intensity of mouth motion can be estimated by averaging the optical flow of all pixels in the mouth region:

$$\overline{O} = \frac{\sum_{(x,y)\in \mathbf{R}} ||\boldsymbol{o}(x,y)||_2}{\#(\mathbf{R})}, \qquad (4)$$

where o(·) is the optical flow vector of each pixel. We further compute the orientations of mouth motion, also based on optical flow. Given the vectors of optical flow at mouth region R, the orientations of mouth motion can be represented by the following histogram:

$$\text{hist}_l = \frac{\sum_{(x,y)\in \mathbf{R}} ||\boldsymbol{o}_l(x,y)||_2}{\#(\mathbf{R})}, \quad l = 1, 2, \ldots, L. \qquad (5)$$

In (5), ol(·) is the orientations of optical flow belonging to the l-th orientation. There are L equal bins for the orientation histogram of (5), i.e., the bin width is 360◦/L. In this paper, we set L to be 8, corresponding to 8 directions of mouth movement. Finally, SVM with the radial bias function (RBF) kernel is used in our algorithm to train the binary classifier for speaking and non-speaking. The input feature vector of SVM consists of mouth motion intensity O, mouth motion orientation histogram [hist1,..., histK], mouth elongation V and average binary value B at three neighboring frames. As a result, the action of speaking can be detected, as one of the high-level features for our saliency prediction method. C. Detection on Head Pose and Head Turning It has been demonstrated in Observation 4 that visual attention on face is relevant to its pose. We thus need to detect the head pose as a feature for predicting video saliency. In [46], 68 landmarks are detected for frontal face, whereas 39 landmarks are found for profile face. In this paper, we estimate the head pose on the basis of the number of landmarks of the tracked face (by Section IV-A). That is, the face is viewed as a frontal face when it has 68 landmarks; otherwise, it is considered to be a profile face given 39 landmarks. Note that a detected face can only have 68 landmarks (frontal) or 39 landmarks (profile). Observation 7 has pointed out that visual attention is also correlated with head turning. Due to this, we further detect the action of head turning, which has two categories: front-toprofile or profile-to-front. In fact, head turning can be tracked in a straightforward manner according to the change of head pose (defined above). We empirically find that the duration of head turning is normally 1 second. Thus, once a head pose change is detected, the corresponding face of adjacent frames within 1 second is annotated as head turning. V. SALIENCY PREDICTION After extracting the above features, our method introduces the M-HMM model and postprocessing step to generate saliency maps of multiple-face videos. The overall pipeline of our method is summarized in Figure 10. As can be seen in this figure, the input is frames of multiple-face videos, and the output is the corresponding saliency map. After face detection and feature extraction, M-HMM is used to predict the attention

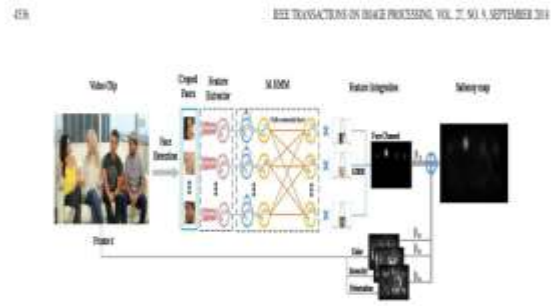


Fig. 10. Pipeline of our proposed method.

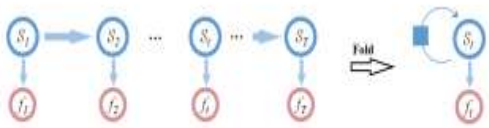


Fig. 11. Overview structure for HMM.

weight of each face by exploring the temporal transition of salient faces across video frames. In our saliency prediction method, we extend HMM to be M-HMM, by allowing more than one interactive state at one time period. Besides, each state in M-HMM depends on the observed features and the previous states. More details about HMM and M-HMM are to be discussed in Sections V-A and V-B, respectively. Finally, a post-processing step is adopted to generate saliency maps of multiple-face videos, as discussed in Section V-C. A. HMM for Single-Face Saliency First, we concentrate on the

application of HMM in our saliency prediction method. Figure 11 shows the structure of HMM. In HMM, we treat high-level static/dynamic feature ft (discussed in Section IV) as the observed feature at the t-th frame. State St , the sequential unit in HMM, stands for the variation of saliency attended to one face. In our application, we have St ∈ {+δ1, 0, −δ2}, where δ1 (> 0) and δ2 (> 0) define the amounts that saliency increase and decrease for a face. Moreover, St = 0 indicates that the saliency of the face remains unchanged across frames. In HMM, the value of the currently processed state St relies on its previous state St−1 and observed feature ft . As such, the saliency map of a video frame is determined by its observed high-level features and the saliency of the face at the previous frame. However, HMM can only deal with one face, since there is one state in each time period for HMM. In the next subsection, we present our M-HMM algorithm to predict the saliency of more than one face. B. M-HMM for Multiple-Face Saliency For M-HMM, multiple HMMs are adopted and combined, each of which is in accordance with the saliency of one face. Figure 12 shows the structure of our M-HMM, in which there are N states in total for a time period. In our saliency prediction method, each state (among N states) means saliency variation of one face at the t-th frame, and they are denoted as {S(n) t }N n=1. Consequently, M-HMM can be applied to the

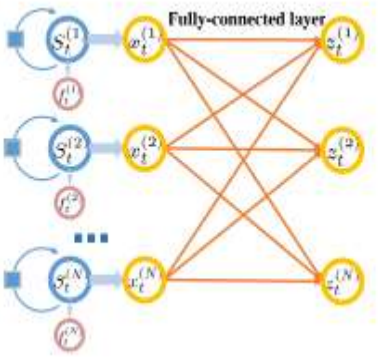


Fig. 12. Structure of M-HMM. Note that the fully-connected layer is different from that in deep learning, because no weight needs to be learnt in this layer.

multiple-face scenarios. As with S_t, the possible values of $S_t^{(n)}$ are $\in \{+\delta_1, 0, -\delta_2\}$. Then, all N states in M-HMM are simultaneously transited along with the processed video frames. Similar to HMM, the states of $\{S_t^{(n)}\}_{n=1}^N$ depend on their corresponding observations of high-level features $f_t^{(n)}$, as well as their previous states $\{S_{t-1}^{(n)}\}_{n=1}^N$.

In the following, we introduce a fully-connected network in M-HMM, via adopting the basic idea of RNN [52]. *Observation 1* has pointed out that most visual attention is attracted by one face among all faces. In other words, if one face receives a large amount of attention in a video frame, then the other faces normally draw few fixations. That is, saliency maps of different faces are highly correlated with each other in a video frame. Thus, $\{S_t^{(n)}\}_{n=1}^N$ at one time period need to be interactive with each other. Accordingly, our M-HMM algorithm takes into account the interaction of state set $\{S_t^{(n)}\}_{n=1}^N$ by adding a fully-connected network. Specifically, we denote $\{z_t^{(n)}\}_{n=1}^N$ ($\in [0, 1]$) as the set of weights, reflecting the proportions of attention belonging to different faces in a video frame. Additionally, $\{x_t^{(n)}\}_{n=1}^N$ is the intermediate units for computing $\{z_t^{(n)}\}_{n=1}^N$. A higher $x_t^{(n)}$ corresponds to a larger $z_t^{(n)}$. Assuming that $\sum_{n=1}^N z_t^{(n)} = 1$, the following *softmax* activation function is used to formulate weights $\{z_t^{(n)}\}_{n=1}^N$ in M-HMM:

$$z_t^{(n)} = \frac{\exp(x_t^{(n)})}{\sum_{n'=1}^N \exp(x_t^{(n')})}, \tag{6}$$

where $x_t^{(n)}$ is defined as

$$x_t^{(n)} = z_{t-1}^{(n)} + S_t^{(n)}. \tag{7}$$

In (7), $x_t^{(n)}$ of one face is determined by saliency variation $S_t^{(n)}$ (i.e., states of M-HMM) and the weight of face attention $z_{t-1}^{(n)}$ at the previous frame. They are modeled as hidden units of the fully-connected network in our M-HMM structure (as shown in Figure 12).

Finally, M-HMM is able to output weights $\{z_t^{(n)}\}_{n=1}^N$. Given $\{z_t^{(n)}\}_{n=1}^N$, we can make use of the dynamic feature $f_t^{(n)}$ to predict the visual attention on each face at the t-th frame. In this paper, the predicted visual attention of the face channel is modeled using the conspicuity map,[4] denoted as $\mathbf{M}_t^F$. It can be computed by

$$\mathbf{M}_t^F = \sum_{n=1}^{N} z_t^{(n)} c_t^{(n)} \mathbf{M}_t^{F_n}, \tag{8}$$

where $\mathbf{M}_t^{F_n}$ denotes the conspicuity of the n-th face upon feature $f_t^{(n)}$, and $c_t^{(n)}$ is the center-bias weight of each face. In our method, $\mathbf{M}_t^{F_n}$ is calculated by the latest work [22], which models the conspicuity map of a face with the Gaussian mixture model (GMM). It is worth pointing out that in [22] the conspicuity map of each face is proportional to its size, with the relationship learnt from training data. As such, face size is already considered in our method, satisfying *Observation 2* of this paper. In addition, *Observation 3* has revealed that visual attention is also correlated with the center-bias feature of faces in multiple-face videos. Therefore, we follow the way of [53] to take into account the face center-bias feature by weighting the Gaussian model $c_t^{(n)}$ in (8). Assuming that $d_t^{(n)}$ is the Euclidean distance of the n-th face to the video center at the t-th video frame, $c_t^{(n)}$ of (8) can be calculated using the following Gaussian model:

$$c_t^{(n)} = \exp\left(-\frac{(d_t^{(n)} - \min_n d_t^{(n)})^2}{\sigma^2}\right). \tag{9}$$

In (9), σ is the standard deviation of the Gaussian model, which reflects the degree of center-bias. Note that Gaussian center-bias weights of (9) are only imposed on conspicuity of each face in our method, rather than all pixels as in [53]. Now, the remaining task is to learn the parameters of our M-HMM for estimating z (n) t , such that the conspicuity of each face can be yielded by (8). At the beginning, all initial states S(n) 1 are simply set to 0 for M-HMM. Next, the matrices of transition probabilities and emission probabilities are two important parameters of M-HMM to be learnt. In our M-HMM, the matrices of these two parameters are identical across different HMMs. It is because transition probabilities and emission probabilities of each HMM are independent of other HMMs, as can be seen in Figure 12. In our method, we apply the maximum likelihood estimation [54] to learn these two matrices from training data. Given the learnt matrices, the Viterbi algorithm [55] is adopted to perform the transition between the previous state and the current state, based on the observed dynamic feature f (n) t of each face.

C. Feature Integration

According to *Observations 6 and 7*, the high-level features $f_t^{(n)}$ can be the actions of speaking and head turning, for predicting video saliency. Accordingly, we define the set of the high-level dynamic features as $\{f_{t,k}^{(n)}\}_{k=1}^{K}$. Specifically, $f_{t,1}^{(n)} \in \{1, 0\}$ means whether the face speaks $(= 1)$ or does not $(= 0\text{d})$. $f_{t,2}^{(n)} \in \{1, 0\}$ indicates whether the head turns from front to profile, and $f_{t,3}^{(n)} \in \{1, 0\}$ indicates whether the face has the profile-to-front turning. Besides, since *Observation 4* has shown that the frontal face receives more attention than the profile face, we further include the static feature of head pose $f_{t,4}^{(n)}$, which stands for frontal face $(= 1)$ or profile face $(= 0)$. At the k-th frame, we can generate the set of face conspicuity maps $\{\mathbf{M}_{t,k}^{F}\}_{k=1}^{4}$, corresponding to different features $\{f_{t,k}^{(n)}\}_{k=1}^{4}$.

Then, we need to combine all conspicuity maps of $\{\mathbf{M}_{t,k}^{F}\}_{k=1}^{4}$ for predicting the face saliency of multiple-face videos. Let $\mathbf{S}_t^F$ be the face saliency of the t-th video frame. It can be computed by the linear combination:

$$\mathbf{S}_t^F = \sum_{k=1}^{4} w_k \mathbf{M}_{t,k}^{F}, \tag{10}$$

where w_k is the weight of the k-th conspicuity map.

Finally, we can compute (10) to predict saliency maps of multiple faces in a video, once the values of $\{w_k\}_{k=1}^{4}$ are known. In fact, the weights of w_k can be learnt from training data via solving the following optimization formulation:

$$\begin{aligned} &\arg \min_{\{w_k\}_{k=1}^{4}} \sum_{l=1}^{L} \| \sum_{k=1}^{4} w_k \mathbf{M}_{l,k}^{F*} - \mathbf{S}_l^{F*} \|_2, \\ &s.t. \sum_{k=1}^{4} w_k = 1, \quad w_k = 1 > 0, \end{aligned} \tag{11}$$

where $\{\mathbf{M}_{l,k}^{F*}\}_{l=1}^{L}$ are the conspicuity maps and $\{\mathbf{S}_l^{F*}\}_{l=1}^{L}$ are human fixation maps, for all L training video frames. In this paper, we apply the disciplined convex programming (CVX) to solve the above optimization formulation.

In order to consider both low-level and high-level features in saliency prediction, our method combines face saliency $\mathbf{S}_t^F$ with saliency maps of three low-level features of GBVS [13] (i.e., $\mathbf{S}_t^I$ for intensity, $\mathbf{S}_t^C$ for color and $\mathbf{S}_t^O$ for orientation). In addition, the weights for the linear combination are determined through the least square fitting on training data. Afterwards, the final saliency map $\mathbf{S}_t$ of each video frame can be yielded for multiple-face videos.

VI. Model Evaluation

A. Setting

In our experiments, we tested all 65 videos in our eye-tracking database (mentioned in Section II). Here, 5-fold cross validation was applied, in which 65 videos were equally divided into 5 non-overlapping sets. One set was used for the test with the others being training sets. Following this way, all 5 sets can be tested. In this paper, the saliency prediction results are reported by averaging over all 65 videos in 5-fold cross validation. Note that both speaking detection

and saliency prediction were trained and tested with the same 5-fold cross validation. Besides, we simply utilized the face detector and head pose detector provided by [46], which had been already trained over the external data of [46].

For speaking detection, the threshold of binary process on gray scale mouth was empirically to be $th_G = 28$, in our experiments. Furthermore, the SVM (with the RBF kernel) of the LIBSVM toolbox [56] was applied, which detects speaking actions of all test videos in 5-fold cross validation. In the LIBSVM toolbox, the penalty parameter C and kernel parameter γ were tuned by grid search on training data. Specifically, the grid search was divided into two steps: one for loose grid search on $C = 2^{-5}, 2^{-4}, ..., 2^{9}$ and $\gamma = 2^{-15}, 2^{-14}, ..., 2^{9}$ (the optimal results are $C = 2^{3}$ and $g = 2^{5}$), and then the other for a fine grid search on $C = 2^{2}, 2^{2.2}, ..., 2^{4}$ and $\gamma = 2^{4}, 2^{4.2}, ..., 2^{6}$. The final optimized parameters were $C = 6.96$ (i.e., $2^{2.8}$) and $\gamma = 18.38$ (i.e., $2^{4.2}$) for our experiments.

For saliency prediction, the values of latent state $S^n_{l,k}$ in M-HMM were tuned to be $\delta_1 = \delta_2 = 0.38$. When training the matrices of the transition and emission probabilities for M-HMM, the values of $z_l^{(n)}$ were obtained by computing the proportion of human fixations on the n-th face to fixations on all faces. When training the weight of each high-level feature channel in (11), the fixations on face regions in the training frames were smoothed with a two-dimensional Gaussian filter (with the cut-off frequency being 6 dB) to obtain $\{\mathbf{S}_l^{F*}\}_{l=1}^{L}$. In addition, all fixations of each training frame were smoothed with the same Gaussian filter, to train the weights of channels on face and low-level features.

B. Evaluation on Feature Detection

In this section, the extraction of high-level features is evaluated, as it is the foundation of saliency prediction. First, we evaluate the performance of our speaking detection algorithm proposed in Section IV-B. Recall that the manually annotated speaking results are available in our eye-tracking database (https://github.com/yufanLiu/find), and they are considered to be the ground truth for speaking detection. The state-of-the-art of speaking detection algorithms [51] and [50] were compared with our algorithm. The metrics of F-measure, accuracy, false positive rate (P_{FP}) and false negative rate (P_{FN}) are measured for evaluation. Here, F-measure is calculated as follows,

$$F_1 = \frac{2P_{TP}}{2P_{TP} + P_{FP} + P_{FN}}, \tag{12}$$

where P_{TP} represents the true positive rate. Note that accuracy is the ratio of correctly detected speaking and non-speaking faces to the total number of faces, at all frames of test videos. Table II reports the results of the three algorithms for all test videos in 5-fold validation. It can be seen that our speaking detection algorithm is significantly superior to [50] and [51], in terms of overall performance measured by F-measure and accuracy. Although [50] has the smallest false negative rate ($P_{FN} = 0.06$), its false positive rate is extremely high ($P_{FP} = 0.84$). By contrast, our algorithm achieves the best false positive rate ($P_{FP} = 0.13$) and its false negative rate ($P_{FN} = 0.38$) is comparable to that of other algorithms. In other words, our algorithm performs the best among all three algorithms on speaking detection. Note that our database is tough for speaking detection because there are multiple faces in the videos and some of them are small, blurry and partially occluded. Meanwhile, our speaking detection algorithm relies on the face alignment algorithm [46] to handle occlusion, pose changes and illumination.

TABLE II
EVALUATION ON SPEAKING DETECTION BY OUR AND OTHER TWO COMPARATIVE ALGORITHMS

	F-measure	Accuracy	P_{FN}	P_{FP}
Our	**0.63**	**0.80**	0.38	**0.13**
[51]	0.35	0.35	0.36	0.76
[50]	0.45	0.37	**0.06**	0.84

Moreover, we show the effectiveness of our detection method on head pose and head turning. For head pose detection, we found from our experiments that its accuracy is approximately 99.1%, averaged over all test videos, which is close to the 99.9% accuracy reported in [46]. For head turning detection, the average accuracy is 90.1%, which is similar to the accuracy of head pose detection as head turning is based on the results of detected head pose. In a word, head pose and head turning can be effectively detected in our method.

C. Evaluation on Saliency Prediction

In this section, we compare our method with 8 conventional saliency prediction methods, including Xu *et al.* [22], Jiang *et al.* [23], SALICON [19], GBVS [13], Rudoy *et al.* [29], PQFT [16], Surprise [9] and OBDL [12]. Additionaly, [13], [19], [22], and [23] are image saliency prediction methods. To be more specific, [22] and [23] work on face saliency prediction of images, which incorporate the high-level static features of face. We compare our method to these two high-level based methods, as there is no face saliency prediction method for videos. On the contrary, [19] is a state-of-the-art deep neural network (DNN) method that automatically learns hierarchical static features for saliency prediction. Besides, [13] is a low-level based method, which provides the saliency of low-level features for our method. Therefore, [13] and [19] are also included in our comparison.

Note that we use our multiple-face tracking technique to detect faces for [22], since it only handles the single-face scenario.

The most recent work of [57] and [58] reported that normalized scanpath saliency (NSS) and correlation coefficient (CC) perform the best among all metrics in evaluating saliency prediction accuracy.[5] Thus, we compare our method with 8 other methods in terms of NSS and CC. Table III reports the comparison results of saliency prediction, averaged over all test videos in the 5-fold cross validation. As shown in this table that our method is much better than all other methods in predicting the saliency of multiple-face videos. Specifically, our method significantly outperforms all video saliency

[5] [57] also showed that area under ROC (AUC) is the worst metric in measuring the accuracy of saliency prediction.

TABLE III
ACCURACY OF SALIENCY PREDICTION BY OUR METHOD AND 8 OTHER METHODS, AVERAGED OVER ALL TEST VIDEOS IN THE 5-FOLD CROSS VALIDATION

	Our	Our+manual	Xu *et al.* [22]	SALICON [19]	Jiang *et al.* [23]	GBVS [13]	Rudoy *et al.* [29]	PQFT [16]	Surprise [9]	OBDL [12]
NSS	**3.61**	3.88	3.14	2.96	0.97	1.23	1.42	0.88	0.88	1.62
CC	**0.66**	0.72	0.61	0.52	0.29	0.33	0.36	0.22	0.21	0.30

TABLE IV
SHUFFLED AUC OF SALIENCY PREDICTION BY OUR METHOD AND 3 OTHER STATE-OF-THE-ART METHODS

	Our	Xu *et al.* [22]	Jiang *et al.* [23]	GBVS [13]
Shuffled AUC	**0.61**	0.58	0.44	0.53

prediction methods in both NSS and CC. Moreover, our method performs much better than the latest DNN method SALICON, with 0.65 and 0.14 increases in NSS and CC, respectively. Furthermore, our method has 0.47 and 0.05 improvements in NSS and CC compared with [22]. These improvements are due to the following reason: The saliencies of all faces have equal importance in [22], whereas the use of high-level dynamic features enables our method to precisely predict salient faces across frames. Moreover, note that both our method and [22] are superior to [23], which imposes unequal importance

on different faces in an image. The main reasons are as follows: (1) The predicted saliency of [23] suffers from incorrectly detected faces because it is based on image face alignment [46], and (2) the utilization of highlevel static features in [23] may predict incorrect salient faces in a video. Conversely, the high-level dynamic features of our method are highly effective in finding the salient faces in a video.

Since the above comparison takes into account the influence of center-bias embedded in saliency prediction methods, we further compare the saliency prediction performance in terms of shuffled AUC, which removes the influence of center-bias. Table IV reports the shuffled AUC results of our method and Xu et al. [22], Jiang et al. [23] and GBVS [13] methods, which bias the saliency prediction toward the center. It can be seen that our method still performs better than the other methods, when removing the influence of center-bias in saliency prediction. In Section VI-D, we further analyze the influence of center-bias in our saliency prediction method in more detail.

Next, we move to the comparison of subjective results. We show in Figure 13 the saliency maps of several frames in a video, generated by our method and 8 other methods. As shown in this figure, our method is capable of finding the salient face according to high-level dynamic features. Consequently, the saliency maps of our method are more accurate than those of other methods. For example, we can see from Figure 13 that the face of the girl is much more salient than the other, when she is speaking (the first column) or turning her head (the last column). Moreover, the man's face is more salient, when he is speaking (the second and third columns) or the girl's face is profile. In contrast, [22] finds all three faces as salient ones, and [23] misses the salient

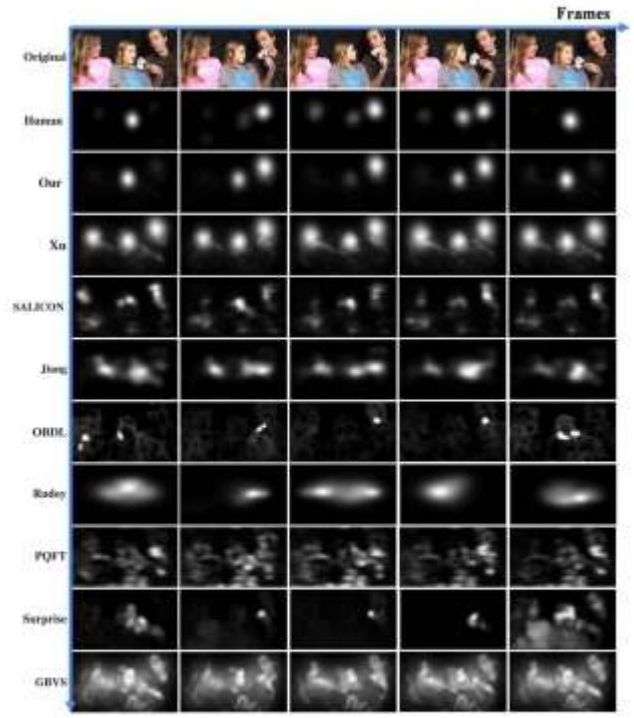


Fig. 13. Saliency maps for different frames of a video selected from our database. These maps are generated by ground truth human fixations, our method, Xu *et al.* [22], SALICON [19], Jiang *et al.* [23], OBDL [12], Rudoy *et al.* [29], PQFT [16], Surprise [9] and GBVS [13].

face of the speaking man because he is far from the video center. In addition, although the predicted saliency of [19] involves some detected faces benefiting from the learned features of DNN, it fails to

predict the transition of the salient face. It is mainly because [19] focuses on image saliency prediction, without considering temporal information or highlevel dynamic features. Figure 14 provides the saliency maps of the frames selected from 5 videos. It is worth pointing out that in the fourth video of Figure 14, all 9 faces are singing simultaneously. In this case, people usually look at each singer, and then concentrate on the singer located in the center. Fortunately, Figure 14 shows that our method can successfully detect the salient face, benefiting from the incorporated centerbias feature. Similarly, the last column of Figure 14 further shows that our method is able to locate the salient face by taking advantage of the center-bias feature, when one face is speaking and some of the other faces are acting. We can further see from the fourth column of Figure 14, our method can still find the salient face when more than one face speaking, benefitting from other features (e.g., the center-bias feature). Again, this figure verifies that our method is able to precisely locate salient faces by turning from actions to saliency.

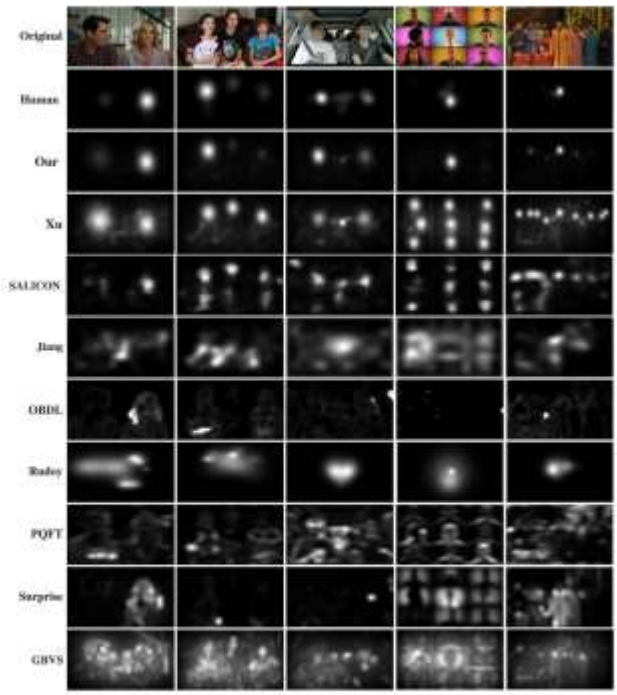


Fig. 14. Saliency maps for several frames selected from different videos in our database. These maps are generated by ground truth human fixations, our method, Xu *et al.* [22], SALICON [19], Jiang *et al.* [23], OBDL [12], Rudoy *et al.* [29], PQFT [16], Surprise [9] and GBVS [13].

D. Performance Analysis of Saliency Prediction Section VI-C has validated that the high-level dynamic features are rather effective in improving the performance of saliency prediction for multiple-face videos. However, these features are automatically detected by the technique of Section IV, which may incur some detection errors as verified in Section VI-B. Thus, it is interesting to see the influence of the feature detection errors on saliency prediction. In Table III, we present the NSS and CC of our method with manually annotated dynamic features. We find that there is a 0.27 NSS improvement or a 0.06 CC improvement, when using manual annotation instead of automatic annotation on high-level features. Thus, the performance of our method can be further improved, via advancing the technique of feature extraction.

Next, we analyze the performance of each individual feature and the feature integration in our method. Figure 15 plots the NSS and CC of our method with each single feature and with all features integrated together. Additionally, the results of [22] are also provided, since our method weights the detected salient faces of [22] with respect to several proposed features. Obviously, we can see from Figure 15 that all single features perform better than [22], validating the effectiveness of each single feature in our method. Besides, one may observe that the feature of speaking is more effective than the features of head turning and head pose in predicting video saliency. More importantly, Figure 15 shows that the integration of all high-level features is superior to each single feature in saliency prediction. This verifies the effectiveness of the feature integration in our method.

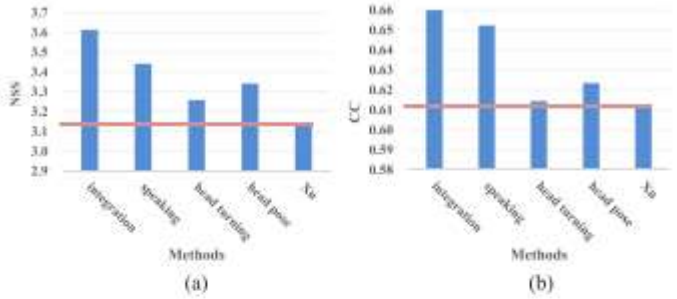


Fig. 15. Performance comparison of our method with different features and the method of [22].

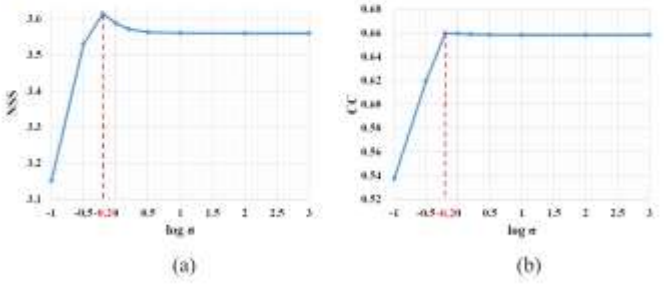


Fig. 16. Saliency prediction performance versus different center-bias parameter σ of (9).

Finally, it is necessary to investigate the effectiveness of the feature of face center-bias in our method. To this end, standard deviation σ in (9) is traversed, imposing different impact of face center-bias on saliency prediction. Figure 16 plots the NSS and CC results at different σ, averaged over all training videos of the 5-fold cross validation. It is clear that the best performance is achieved once $\sigma = 10^{-0.2}$, and thus, σ was set to $10^{-0.2}$ in our above experiments. This figure also shows that when σ increases from $10^{-0.2}$ to 10^{3}, the accuracy of saliency prediction slightly decreases. This implies that the feature of face center-bias is effective in our method, since its impact dramatically decreases in (9) for $\sigma = 10^{-0.2} \rightarrow 10^{3}$. On the other hand, a small value of σ causes our method to only predict the face closest to the video center as the most salient one, according to (8) and (9). When σ is as small as 10^{-1}, the NSS of our method is ~ 3.16, reflecting the performance of the single feature of face center-bias.

For time complexity, our method consumes roughly 2.37 seconds per frame. Our method was implemented in Matlab R2016b and run on a computer with a Intel Core i7-6700K CPU@4.00 GHz and RAM 32.0GB. Specifically, the time consumption of our method includes face detection and landmark localization (2 s per frame), face tracking (287 ms per frame), feature extraction (52 ms per frame), M-HMM (0.6 ms per frame) and feature integration (40 ms per frame). To improve the speed of our method, some fast algorithms of face detection and landmark localization may be applied, e.g., [48].

VII. IMPLEMENTATION IN VIDEO COMPRESSION

The proposed saliency prediction method has potential to be implemented in some tasks of video processing. For instance, in human-centered multimedia, our method may be utilized to locate salient faces in a video, seen as ROI. Then, the quality of experience (QoE) of video conferencing can be improved by assigning more coding bits to salient faces, during video compression. In this section, we present a simple implementation of our saliency prediction method in the compression of video conferencing, which is embedded into the latest HEVC standard.

A. Method for Video Compression

When encoding a multiple-face video frame by HEVC, our implementation allocates target bits to each coding tree unit (CTU) according to the video saliency predicted by our method. Specifically, our implementation is embedded into the r-λ rate control (RC) scheme [59] of HEVC. In the conventional HEVC, the RC scheme [59] estimates the bit per pixel (bpp) at each CTU given a target bit-rate, for rate-distortion optimization. Instead, we follow our previous work [60] to define bit per saliency weight (bpw), for perceptual rate-distortion optimization (also called the perceptual RC scheme) in HEVC. For the t-th frame, assuming that $\text{bpw}_{t,i}$ is the bpw of the i-th pixel, the target bit $r_{t,j}$ for the j-th CTU can be determined by

$$r_{t,j} = \sum_{i \in \mathbf{I}_{t,j}} \text{bpw}_{t,i}, \tag{13}$$

where $\mathbf{I}_{t,j}$ is the set of pixels in the j-th CTU. Before encoding a frame of a multiple-face video, $\text{bpw}_{t,i}$ in (13) can be obtained from the saliency map $\mathbf{S}_t$ generated by our saliency prediction method. Let $\mathbf{S}_t(i)$ be the predicted saliency value of the i-th pixel at the t-th frame. Then, we have

$$\text{bpw}_{t,i} = \frac{\mathbf{S}_t(i) \cdot r_t}{\sum_{i \in \mathbf{I}_t} \mathbf{S}_t(i)}, \tag{14}$$

where r_t and $\mathbf{I}_t$ are the target bit-rate and pixel number of the t-th frame, respectively.

Next, the average bpw in each CTU can be estimated by

$$\overline{\text{bpw}}_{t,j} = \frac{r_{t,j}}{\#(\mathbf{I}_{t,j})}. \tag{15}$$

where $\#(\mathbf{I}_{t,j})$ indicates the overall number of pixels in the j-th CTU. Then, we make $\overline{\text{bpw}}_{t,j}$ instead of average bpp in the conventional RC scheme [59], such that the following exists for perceptual RC in HEVC:

$$\begin{aligned} \lambda_{t,j} &= \alpha_{t,j} \cdot (\overline{\text{bpw}}_{t,j})^{\beta_{t,j}}, \\ QP_{t,j} &= c_1 \cdot ln(\lambda_{t,j}) + c_2, \end{aligned} \tag{16}$$

In (16), for each LCU, $\lambda_{t,j}$ is the Lagrange multiplier of optimization, and $QP_{t,j}$ is the quantization parameter (QP) as the output of RC. In addition, $\alpha_{t,j}$ and $\beta_{t,j}$ are the parameters to estimate the r-λ relationship; c_1 and c_2 are the fitting parameters for QP estimation. Refer to [59] for more details on how to update these parameters alongside compressed frames. Finally, each frame of video conferencing can be encoded by HEVC, on the premise of the CTU-wise QPs estimated by our perceptual RC. Figure 17 summarizes the overall procedure of our implementation in perceptual RC for HEVC-based video compression.

Fig. 17. The framework of our perceptual RC on the basis of our saliency prediction method.

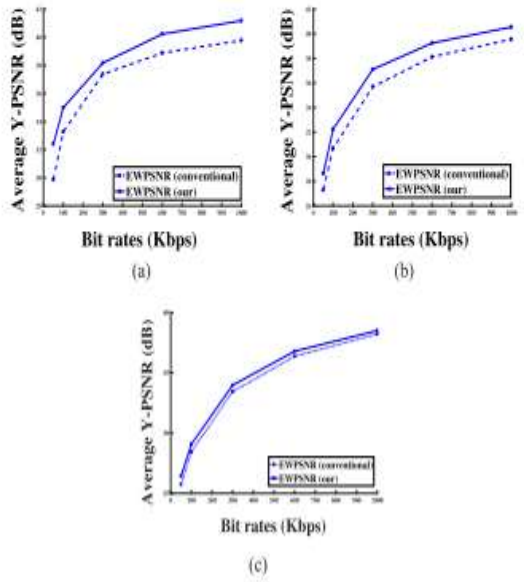


Fig. 18. Rate-distortion curves of our and conventional schemes.

Results of Video Compression In this section, we report the compression results to validate the performance of the above implementation. Since our saliency prediction method is capable of locating salient faces in a multiple-face video, our experiments test on the sequences of Class E (the class of video conferencing) from the JCTVC database [61]. In the JCT-VC database, Class E consists of three 720p raw sequences: Johnny, KristenAndSara and FourPeople. The HEVC reference software HM 16.0 (in the LowDelay configuration) was used to compress all those three sequences at different bit-rates, with its conventional r-λ [59] and our perceptual RC schemes.

Here, the eye-tracking weight PSNR (EWPSNR) [62] is used to evaluate the distortion of compressed sequences at various bit-rates. Note that EWPSNR weights PSNR with human fixation maps, thereby well reflecting the subjective quality of compressed sequences. Figure 18 plots the ratedistortion curves of compressing all three test sequences, in terms of EWPSNR. We can see from this figure that EWPSNR of our perceptual RC implementation is much better than the conventional HEVC compression, with approximately 1-2 dB improvement. Thus, we can conclude that the implementation of our saliency prediction method is able to improve the perceptual quality of HEVC compression on video conferencing.

Fig. 19. Subjective quality comparison of KristenAndSara. (a) and (b) are the 720th frame compressed at 300 Kbps by the conventional and our scheme, respectively.

TABLE V
DMOS COMPARISON OF HEVC AND OUR APPROACH AT 300 Kbps BIT RATE

Test Sequence	DMOS (conventional)	DMOS (our)	DMOS difference
Johnny	32.08	28.99	-3.09
KritenAndSara	45.23	36.14	-9.09
FourPeople	45.72	38.05	-7.67

We further compare the subjective quality of our implementation and conventional compression in Figure 19. One may observe from this figure that our implementation yields more satisfactory quality in ROI (the salient face) with some quality loss in non-ROI, compared to the conventional HEVC compression. To quantify the subjective quality, we conducted the difference mean opinion score (DMOS) experiment using the single stimulus continuous quality scale (SSCQS) procedure of Rec. ITU-R BT.500 [63]. In the DMOS experiment, 12 subjects were asked to rate the quality of sequences compressed at 300 Kbps. The quality rate scales are divided as: excellent (100-81), good (80-61), fair (60-41), poor (40-21) and bad (20-1). Since DMOS measures the difference of the rated scores between uncompressed and compressed sequences, smaller DMOS indicates better subjective quality. Table V tabulates the DMOS results of our and conventional HEVC compression. As shown, the subjective quality of our perceptual RC is superior to the conventional one. This again verifies the potential implementation of our saliency prediction method in video compression.

VIII. CONCLUSION

In this paper, we have proposed a novel saliency prediction method for multiple-face videos, which learns to predict the salient face with regard to some static and dynamic highlevel features of faces. First, we established an eye-tracking database consisting of 65 multiple-face videos. Then, we found out from our database that visual attention in multiple-face videos is highly correlated with both static and dynamic features of face at high-level. These features include face size, center-bias, speaking, head turning and head pose. Accordingly, we developed the techniques to extract these features. Next, a new M-HMM algorithm was proposed to integrate the observed features and saliency transition from previous frames into a uniform framework. This way, the high-level features, such as actions of speaking and head turning, can be turned to video saliency, for predicting who to look at. The experimental results demonstrated that our method is able to advance state-of-the-art saliency prediction on multiple-face videos. Finally, we provided a potential implementation of our saliency prediction method in video compression. There exist three directions for the future work. (1) Our database and analysis at the current stage may be lacking generalization, as it mainly handles limited high-level features, e.g., speaking, head turning, and so forth. In the future, the database can be extended to include more general scenarios, and some other high-level features, such as gesture and expression, may be incorporated into the saliency prediction framework.

(2) There is still room to improve saliency prediction accuracy by refining the speech detection algorithm. For instance, the audio component of videos may be taken into account in the speaking detector for saliency prediction.

(3) RNN is an efficient deep learning approach, which shares a similar sequential structure with the proposed M-HMM. Thus, applying RNN to saliency prediction is another promising future work.

(4) Our method only focuses on the visual cues to predict saliency of video. Actually, audio may also have impact on visual attention. Therefore, it is an interesting future work to consider the audio cues in saliency prediction of multipleface videos.

REFERENCES

[1] G. T. Buswell, How People Look at Pictures: A Study of the Psychology and Perception in Art. Oxford, U.K.: Univ. Chicago Press. 1935.

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 185–207, Jan. 2013.

[3] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in Proc. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 2751–2758.

[4] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," ACM Trans. Graph., vol. 29, no. 5, pp. 160:1–160:10, 2010.

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

[6] M. Xu, X. Deng, S. Li, and Z. Wang, "Region-of-interest based conversational HEVC coding with hierarchical perception model of face," IEEE J. Sel. Topics Signal Process., vol. 8, no. 3, pp. 475–489, Jun. 2014.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[8] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," IEEE Trans. Image Process., vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[9] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," Vis. Res., vol. 49, no. 10, pp. 1295–1306, Jun. 2009.

