Learning Dynamic GMM for Attention Distribution on Single-face Videos

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

Abstract

The past decade has witnessed the popularity of video conferencing, such as FaceTime and Skype. In video conferencing, almost every frame has a human face. Hence, it is necessary to predict attention on face videos by saliency detection, as saliency can be used as a guidance of region-of-interest (ROI) for the content-based applications. To this end, this paper proposes a novel approach for saliency detection in single-face videos. From the data-driven perspective, we first establish an eye tracking database which contains fixations of 70 single-face videos viewed by 40 subjects. Through analysis on our database, we investigate that most attention is attracted by face in videos, and that attention distribution within a face varies with regard to face size and mouth movement. Inspired by the previous work which applies Gaussian mixture model (GMM) for face saliency detection in still images, we propose to model visual attention on face region for videos by dynamic GMM (DGMM), the variation of which relies on face size, mouth movement and facial landmarks. Then, we develop a long short-term memory (LSTM) neural network in estimating DGMM for saliency detection of single-face videos, so called LSTM-DGMM. Finally, the experimental results show that our approach outperforms other state-of-the-art approaches in saliency detection of single-face videos.

1. Introduction

Visual saliency [5] aims at predicting how much each pixel or region of an image/video attracts human's attention, which has been widely used in areas of object detection [16], video quality assessment [39] and perceptual video coding [35]. The studies on visual saliency can be traced back to 1998, when Itti and Koch [19] explored 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 [18]. During the past two decades, extensive approaches, such as [17, 4, 38, 9, 30, 24, 13, 8, 12], have been proposed for detecting saliency in videos. All these saliency detection approaches are heuristic ones, as they are generally driven by incorporating biologically-inspired features. However, the biologically-inspired features of these approaches rely heavily on the unmatured study of the human visual system (HVS), leading to inferior performance in saliency detection.

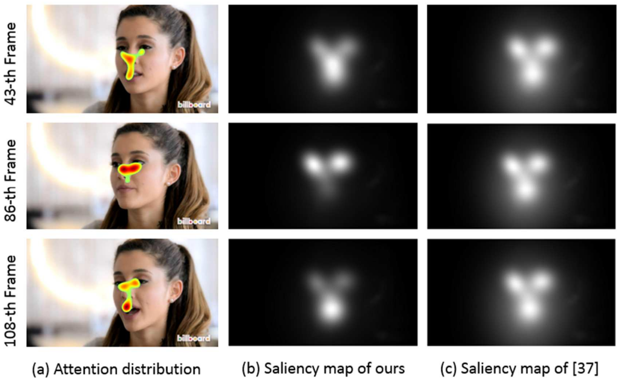
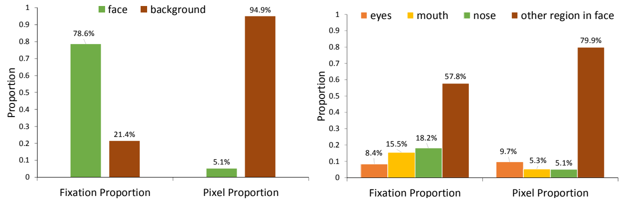


Figure 1: An example of video saliency maps generated by our approach and [37]. Note that [37] is a saliency detection approach for images, while ours works on videos. Here, the saliency maps of [37] are generated by regarding each video frame as a still image. The visual attention distribution by 40 subjects is also shown in this figure.

Recently, the top-down approaches ([21, 14, 15, 27, 31, 7, 22, 28, 36, 6, 40, 20, 37]) have become more prevalent in both image and video saliency detection, which learn saliency model from human fixations on training images/videos. These top-down approaches found out 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 [6, 40, 20, 37]. To be more specific, Cerf *et al.* [6] investigated from eye tracking data that face is highly correlated with attention, and they therefore proposed to integrate face channel with the channels of Itti's model [19] using equal weights, for detecting saliency of face images. Later, Zhao *et al.* [40] found that the face channel is more important than other channels. Accordingly, they proposed to learn the optimal weights of different channels by least square fitting on eye tracking data, further improving the saliency detection performance of [6]. Most recently, Xu *et al.* [37] proposed to model the saliency distribution of face region by Gaussian mixture model (GMM) [3], which is learned from the training data using the conventional expectation maximization (EM) algorithm. All above approaches handle images with face, significantly advancing the development of top-down saliency detection of images.

In contrast to top-down image saliency detection methods, most of the existing video saliency detection approaches [18, 17, 9, 8, 12] make use of the bottom-up information, like motion vector, flicker, as well as spatial and temporal correlation. On the other hand, face videos [25] have undergone explosive growth, due to the emerging video conferencing applications, like FaceTime and Skype. As analyzed in this paper later, face receives more visual attention in videos (77.7% fixations) than that in images (62.3% fixations). Thus, face also plays a vital role in predicting saliency of face videos, similar to its important role in saliency detection of images. The most recent work of [1] has been proposed to predict which face is salient among multiple faces, for multiple-face video saliency detection. However, few work has been devoted to precisely modeling attention distribution within face region for videos. Although videos are composed of images, they are fundamentally viewed differently by people from still images. It is because the dynamic changes of pictures in videos can be also seen as saliency cues. Thus, video saliency cannot be precisely detected merely by the assembly of image saliency, as shown in Figure 1. Figure 1 further shows that face saliency can be modeled as dynamic GMM (DGMM) in videos, in which the GMM distribution of attention varies across video frames.

By establishing an eye tracking database of single-face videos, we find in this paper that most attention is attracted by face in videos, the distribution of which varies with respect to face size and mouth movement. Upon this finding, we propose a long short-term memory (LSTM) based DGMM (LSTM-DGMM) approach to predict saliency distribution within a face for videos, in which attention on face and facial features is modeled by the distribution of GMM. Different from Xu's static learned GMM [37] for image saliency detection, parameters of GMM are dynamically modified in our LSTM-DGMM approach to model DGMM distribution alongside frames, according to the content in videos. Such dynamic parameters can be learned by our LSTM from training fixations in our approach. As far as we know, LSTM [11] is an advanced recurrent neural network (RNN), which can learn long-term dependencies of sequential data. For saliency detection of single-face videos, we thus utilize LSTM to learn the dependency information of DGMM between frames, for modeling the variation of face saliency distribution in a video. Specifically, in a single-face video, we take the content of previous and current frames (i.e., the face size, mouth movement intensity and facial landmarks) as the input to LSTM, and then predict the parameters of DGMM (i.e., means, variances and weights of gaussian components) for the current frame. As such, saliency distribution within a face region can be modeled by DGMM. Finally, we combine the modeled face saliency with saliency detected by the conventional feature channels [19], to predict attention on single-face videos. The experimental results verify that our LSTM-DGMM approach advances the state-of-the-art saliency detection in single-face videos.



(a) Face and background

(b) Subregions in face

Figure 2: Proportions of fixations and pixel numbers in different regions,counted on all 70 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.

2. Database and analysis

2.1. Database

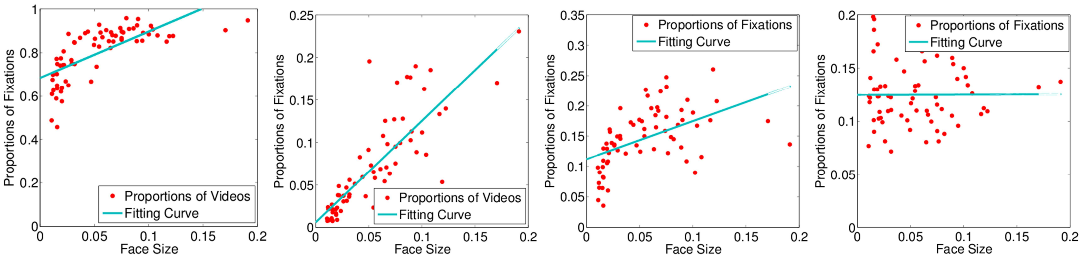
To our best knowledge, there exists no eye tracking database on single-face videos. Therefore, we conducted the eye tracking experiment to obtain a database, in which the fixations on free-viewing single-face videos are available. The database is composed of 70 single-face videos which are selected from 300-VW[33] database and YouTube. The resolutions of all 70 videos in our database are 1280×720, and their frame rates are around 30Hz. There are 40 subjects[1] involved in the experiment to watch all 70 videos, including 24 males and 16 females aging from 21 to 35.

During the experiment, the videos were displayed at their original resolutions (720p), and their display order is 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, $1,006,478$ fixations over $27,707$ frames of 70 videos were collected in our database. Our database would be freely downloadable in the camera-ready version for facilitating the future research.

2.2. Database analysis

We investigate the intrinsic factors which have impact on visual attention to single-face videos, by analyzing the fixa-

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



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

Figure 3: Proportions of fixations on face and facial features versus face sizes, for all 70 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 70 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) 0.66 and (d) 0.07.

tions obtained from 70 videos in our database. We have the following observations. Note that the technique on extracting face-related features for our database analysis is to be discussed in Section 3.

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

First, we show in Figure 2-(a) the proportions of fixations and pixels belonging to face and background, respectively, for all 70 videos. As seen in Figure 2-(a), although the face region only takes up 5.1% pixels in video frames, it attracts 78.6% visual attention. Compared to 62.3% fixations attracted by face in images[2], face region is more salient in drawing visual attention in videos. Besides, Figure 2-(b) illustrates the proportions of pixels and fixations within face region. We can see from this figure that facial features consume 20.1% pixels in face region (9.7% for two eyes, 5.3% for mouth and 5.1% for nose), whereas they draw 42.1% fixations (8.4% for eyes, 15.5% for mouth and 18.2% 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 face, eyes and nose increases along with the enlarged size of face in videos, whereas the attention on mouth is invariant to the face size[3].

Figure 3 shows the proportions of fixations belonging to the regions of face and facial features at different face sizes in 70 videos of our database. In this figure, 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 3, 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 videos. This completes the analysis of Observation 2.

Observation 3: The amount of attention in face region is not affected by mouth movement and blink, whereas its distribution is variant to mouth movement and invariant to blink.

Before figuring out the relationship between visual attention and mouth movement or eye blink, we obtained the ground-truth annotations of eye blink and mouth movement, by manually annotating all 70 videos of our database[4]. Then, the statistical results of fixations versus mouth movement and eye blink are shown in Figure 4, for all 70 videos of our eye tracking database. From this figure, we can find that the proportions of fixations on face are almost the same, whether mouth moves or eyes blink. This implies that the amount of attention on face is invariant to mouth movement and blink. On the other hand, the distribution of attention on face varies with regard to mouth movement. That is, when mouth moves, more attention is drawn to mouth regions with reduced attention on eye regions. Different from mouth movement, the eye blink hardly changes the attention distribution in face. In summary, mouth movement only influences the distribution of attention within face, while attention amount and distribution of face are not sensitive to eye blink. This completes the analysis of Observation 3.

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

Figure 5 plots the fixation proportions of mouth at different intensities for mouth movement, with the scatter analysis on all fixations of mouth regions from our database. Here, the intensity of mouth movement is measured using the method in the next section. We can see from Figure 5 that attention on mouth generally increases with the increased intensity of mouth movement, and the corresponding Spearman rank correlation coefficient is 0.24, much larger than 0.07 for the correlation between fixation proportion in mouth and face size. This completes the analysis of Observation 4.

3. Features extraction

As presented above, face-related features significantly influence the distribution of visual attention on single-face

[2] Note that in [37] face averagely has 5.7% pixels in the whole image.

[3] Here, face size refers to the proportion of pixel number belonging to the face region in a video frame.

[4] There were 3 volunteers to annotate movements of eyes and mouth at all frames in 70 videos. Then, the ground-truth annotations of eyes and mouth movements were obtained by the major voting, which are also available along with our eye tracking database.

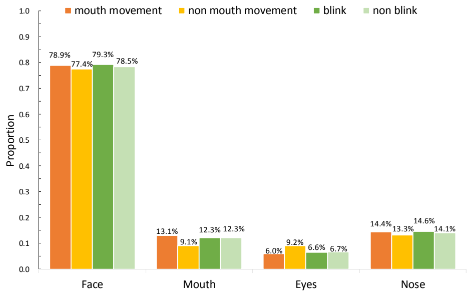


Figure 4: Proportions of fixations belonging to face and facial features averaged on all 70 videos of our database, for the cases of mouth with and without mouth movement as well blink and non-blink.

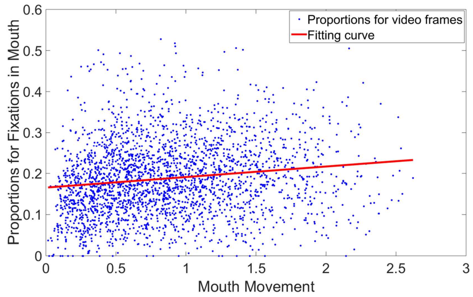
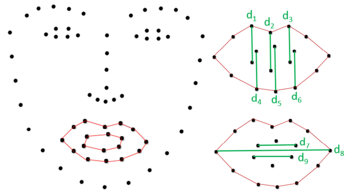


Figure 5: Proportions of fixations in mouth at different intensities for mouth movement. The Spearman rank correlation coefficient here is 0.24.

videos. Thus, this section mainly deals with the extraction of the face-related features of videos, which include face, facial features and mouth movement intensity.

Face and Facial Features. Observation 1 has shown that face and its facial features in a single-face video are much more salient than background. Hence, it is necessary to extract face and facial features for saliency detection. In this paper, we follow the way of [37] to automatically segment the regions of the face and facial features, by leveraging the face alignment algorithm [32]. Specifically, 66 landmark points are located according to point distribution model (PDM) [32]. 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 6-(a) shows an example for the extraction of face and facial features, based on the 66-point PDM.

Intensity of Mouth Movement. Observation 4 has figured out that attention distribution within face is correlated with the movement intensity of mouth. We therefore need to measure the mouth movement intensity. Here, we use the 18 mouth landmarks (see Figure 6) to quantify the mouth movement intensity at the t-th frame (denoted by D_t). Given the mouth landmarks, D_t can be determined upon the difference of width and height between neighboring frames. However, there may be some error on detecting the landmarks of mouth. To reduce the impact of such error on D_t, D_t needs to be calculated by averaging over more than one Euclidean distance, at either horizonal or vertical direction. In our approach, we compute on 9 Euclidean distances: d_1, d_2, d_3, d_4, d_5, d_6 for the vertical distance, and d_7, d_8, d_9 for the horizontal distance. Refer to Figure 6-(b) for more details. Then, the intensity of mouth movement at the t-th frame can be calculated as,



(a) 66-point PDM (b) Mouth

Figure 6: An example of PDM for face and mouth.

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

where $d_i^{(t)}$ and $d_i^{(t-t')}$ are the i-th distances at frames t and $(t-t')$, respectively. In (1), the denominator is used to compute the relative distance for measuring mouth movement intensity. According to the theory of persistence of vision [2], 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 achieved using the calculation of (1) and (2).

4. The proposed approach

4.1. Feature integration

In our approach, we follow the basic way of [37] to integrate saliency maps of different channels together for saliency detection in single-face videos. To be more specific, we assume that $\mathbf{S}_t^C$, $\mathbf{S}_t^I$, $\mathbf{S}_t^O$ and $\mathbf{S}_t^F$ are the conspicuity maps of channels on features of color, intensity, orientation and face at the t-th video frame. Then, these conspicuity maps need to be linearly combined for outputting final saliency map $\mathbf{S}_t$ as follows,

$$\begin{aligned} &\mathbf{S}_t = w_C \mathbf{S}_t^C + w_I \mathbf{S}_t^I + w_O \mathbf{S}_t^O + w_F \mathbf{S}_t^F \\ &\text{s.t.} \quad w_C + w_I + w_O + w_F(s_t) = 1, w_F(s_t) = \sum_{j=0}^{J} a_j {s_t}^j, \end{aligned} \quad (3)$$

where $\mathbf{w} = [w_C, w_I, w_O, w_F(s_t)]^T$ are the weights corresponding to each feature channel, and s_t is the face size at frame t. Based on Observations 2 and 3, the amount of attention in face region is affected by face size, and is invariant to mouth movement. Thus, w_F can be represented by the polynomial function with regard to face size s_t, denoted as $w_F(s_t) = \sum_{j=0}^{J} a_j {s_t}^j$ where $\{a_j\}_{j=0}^{J}$ are the polynomial coefficients to be learned from the training data. In

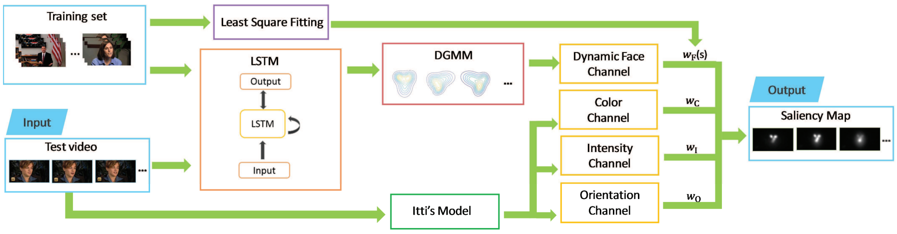


Figure 7: Framework of our LSTM-DGMM approach for saliency detection of single-face videos.

this paper, we adopt Itti's model [19] to yield conspicuity 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, face is incorporated in our approach, as Observation 1 validates that face receives most visual attention in a video. In our approach, we utilize LSTM-DGMM to model conspicuity map $\mathbf{S}_t^F$ of face channel, for saliency detection of single-face videos.

Algorithm 1 The recursive algorithm for solving (7)

```
Initialize the values of $\pi_t(0)$ as $\mathbf{0}$, and $\sigma_t$ as $\sigma_t(0)$ in the way of [37].
Vectorize all the pixel coordinates ($x$ and $y$) in the $t$-th frame into one-dimensional form ($\mathbf{x}$ and $\mathbf{y}$). Similarly, $\mathbf{G}_t$ is vectorized into one dimension $\mathbf{V}(\mathbf{G}_t)$.
for $k = 1$ to $K$ do

  Stage 1. Optimize $\pi_t(k)$ with $\sigma_t(k-1)$:
  for $i = 1$ to 5 do
      $\mathbf{E}_t^i(k-1) = \exp\{-\frac{(\mathbf{x}-\mu_{t,x}^i)^2}{(\sigma_{t,x}^i(k-1))^2} - \frac{(\mathbf{y}-\mu_{t,y}^i)^2}{(\sigma_{t,y}^i(k-1))^2}\}$,
      where $\sigma_{t,x}^i(k-1)$ and $\sigma_{t,y}^i(k-1)$ are defined by (5).
  end for
  Set $\mathbf{E}_t(k-1) = \{\mathbf{E}_t^i(k-1)\}_{i=1}^5$.
  Compute $\pi_t(k)^T = \mathbf{E}_t(k-1)^\dagger \cdot \mathbf{V}(\mathbf{G}_t)$, where $(\cdot)^\dagger$ is the pseudo inverse.
  if $|\mathcal{N}(\mathbf{G}_t, \mathbf{S}_t^F(\sigma_t(k-1), \pi_t(k)) - \mathcal{N}(\mathbf{G}_t, \mathbf{S}_t^F(\sigma_t(k-1), \pi_t(k-1))| \leq \varepsilon$ then
      Break the FOR loop; return $\pi_t(k)$ and $\sigma_t(k-1)$.
  end if

  Stage 2. Optimize $\sigma_t(k)$ with $\pi_t(k)$:
  Calculate $\mathbf{E}_t(k) = \mathbf{V}(\mathbf{G}_t) \cdot (\pi_t(k)^T)^\dagger$.
  for $i = 1$ to 5 do
      Update $\sigma_t(k)$ by
      $(\frac{1}{\sigma_{t,x}^i(k)^2}, \frac{1}{\sigma_{t,y}^i(k)^2})^T = ((\mathbf{x}-\mu_{t,x}^i)^2, (\mathbf{y}-\mu_{t,y}^i)^2)^\dagger \cdot (-\ln \mathbf{E}_t(k)^i)$.
  end for
  if $|\mathcal{N}(\mathbf{G}_t, \mathbf{S}_t^F(\sigma_t(k), \pi_t(k)) - \mathcal{N}(\mathbf{G}_t, \mathbf{S}_t^F(\sigma_t(k-1), \pi_t(k))| \leq \varepsilon$ then
      Break the FOR loop; return $\pi_t(k)$ and $\sigma_t(k)$.
  end if
end for
return $\pi_t(K)$ and $\sigma_t(K)$ as the solution to (7).
```

4.2. DGMM

According to Observation 1, attention on face region is more likely to be drawn by facial features. Thus, GMM of [37] can be applied to model conspicuity map of face. However, different from static GMM for still face images [37], the parameters of GMM should be dynamic across frames in video saliency detection, according to Observation 4. Thus, we propose DGMM in this paper to model the dynamic variation of visual attention on a face. Specifically, for the t-th frame of a single-face video, we can model its conspicuity map $\mathbf{S}_t^F$ as follows,

$$\mathbf{S}_t^F = \sum_{i=1}^{5} \pi_t^i \mathcal{G}_t^i = \pi_t \mathcal{G}_t, \tag{4}$$

where π_t^i is the weight of the i-th GM $\mathcal{G}_t^i$. Here, $\mathcal{G}_t = (\mathcal{G}_t^1, \mathcal{G}_t^2, \mathcal{G}_t^3, \mathcal{G}_t^4, \mathcal{G}_t^5)^T$ and $\pi_t = (\pi_t^1, \pi_t^2, \pi_t^3, \pi_t^4, \pi_t^5)$ correspond to the GMs and their weights, for modeling attention on face, left eye, right eye, nose and mouth. Specifically, for each GM $\mathcal{G}_t^i$, we assume that their mean μ_t^i and standard deviation σ_t^i are

$$\mu_t^i = (\mu_{t,x}^i, \mu_{t,y}^i),$$
$$\sigma_t^i = \begin{pmatrix} \sigma_{t,x}^i & 0 \\ 0 & \sigma_{t,y}^i \end{pmatrix}. \tag{5}$$

In above equation, $\mu_{t,x}^i$ and $\mu_{t,y}^i$ are means of $\mathcal{G}_t^i$ at x and y axes; $\sigma_{t,x}^i$ and $\sigma_{t,y}^i$ are standard deviations. Then, for pixel at (x,y), $\mathcal{G}_t^i$ can be represented by

$$\mathcal{G}_t^i = \exp\{-\frac{(x-\mu_{t,x}^i)^2}{(\sigma_{t,x}^i)^2} - \frac{(y-\mu_{t,y}^i)^2}{(\sigma_{t,y}^i)^2}\}. \tag{6}$$

Once weigh π_t^i, mean μ_t^i and standard deviation σ_t^i are obtained for each GM, conspicuity map of face can be modeled by DGMM of (4) and (6). In this paper, μ_t^i is simply set to be the center of the corresponding face or facial feature, detected by the method of Section 3. As for π_t^i and σ_t^i, we utilize an advanced LSTM network to learn them, to be discussed in Section 4.3.

Before learning LSTM, we need to estimate DGMM distribution upon ground-truth fixations from training data, as the target of LSTM. Since μ_t is known after the detection of face and facial features, the task of DGMM estimation turns to working out π_t and σ_t, formulated by

$$\underset{\pi_t,\sigma_t}{\operatorname{argmax}} \mathcal{N}(\mathbf{G}_t, \mathbf{S}_t^F), \qquad \text{s.t.} \quad \sum_{i=1}^{5} \pi_t^i = 1, \pi_t \geq 0, \tag{7}$$

where $\mathbf{G}_t$ is the distribution of ground-truth fixations and it corresponds to saliency map $\mathbf{S}_t^F$ modeled by DGMM with parameters π_t and σ_t. In addition, $\mathcal{N}(\cdot)$ is the function of normalized scanpath saliency (NSS)[5], which evaluates similarity between human fixation distribution $\mathbf{G}_t$ and saliency

[5] NSS is used here for measuring similarly as [23] proved that NSS is a most effective way in evaluating saliency detection accuracy.

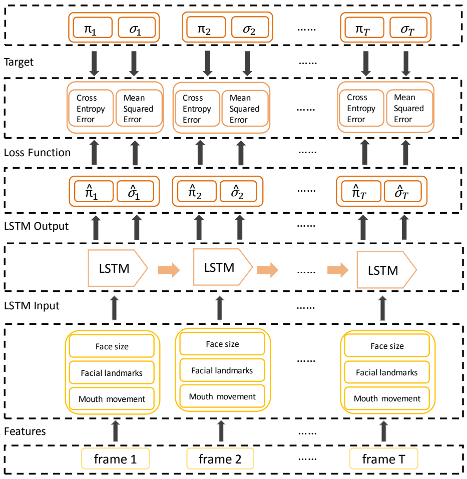

π_1 σ_1 π_2 σ_2 ...... π_T σ_T
Target
Cross Entropy Error
Mean Squared Error
Loss Function
$\hat{\pi}_1$ $\hat{\sigma}_1$ $\hat{\pi}_2$ $\hat{\sigma}_2$ ...... $\hat{\pi}_T$ $\hat{\sigma}_T$
LSTM Output
LSTM
LSTM Input
Face size
Facial landmarks
Mouth movement
Features
frame 1
frame 2
frame T

Figure 8: Structure of our LSTM for DGMM.

maps $\mathbf{S}_t^F$. Note that the greater NSS value means higher similarity. Unfortunately, we find that optimization formulation of (7) is a non-convex problem. In this paper, we solve (7) by developing a recursive algorithm, which iteratively updates π_t and σ_t with the other being fixed. The details about our recursive algorithm are summarized in Algorithm 1.

4.3. LSTM for DGMM

In this section, we present an advanced LSTM network for learning to predict π_t and σ_t, which models DGMM distribution of face saliency in a video. As such, the sequential dependency is considered in the DGMM model by LSTM. However, different from the conventional LSTM [34], our LSTM network is designed for regression problem, as π_t and σ_t are continuous variables. The architecture of our LSTM is shown in Figure 8.

The input to LSTM is the features extracted from each frame of single-face videos. As analyzed in Observations 2 and 4, the DGMM distribution of attention within a face is correlated to face size s_t and the intensity of mouth movement D_t. Thus, s_t and D_t are input to our LSTM. In addition, Observation 1 pointed out that facial features influence the attention distribution within the face region. Due to this, we also take the facial landmarks of facial features as the input to saliency detection. To remove the redundancy of the conventional PDM, 13 landmarks of facial features [29] are selected as the input feature to LSTM, and they are discribed by $\mathbf{l}_t$. This way, the over-fitting can be avoided in training LSTM. Finally, we have feature vector $\mathbf{x}_t = (\mathbf{l}_t, s_t, D_t)$ as the input to LSTM.

Given the input features of face size, facial landmarks and mouth movement, our LSTM incorporates memory cells to learn when to update hidden states with gate on forgetting previous hidden states, for predicting π_t and σ_t. Specifically, for the memory cell $\mathbf{c}$ of LSTM, there are several hidden gates with varying connections. For the memory unit, input gate $\mathbf{g}_i$ controls whether LSTM takes into count its current input $\mathbf{x}_t$, forget gate $\mathbf{g}_f$ decides whether to forget its previous memory $\mathbf{c}_{t-1}$, and the output gate $\mathbf{g}_o$ considers how much to transfer the memory to the hidden state $\mathbf{h}_t$. At the t-th frame, input gate $\mathbf{g}_{i,t}$, forget gate $\mathbf{g}_{f,t}$, and output gate $\mathbf{g}_{o,t}$ are calculated by

$$\begin{aligned} \mathbf{g}_{i,t} &= \sigma(\mathbf{W}_{g_i}\mathbf{x}_t + \mathbf{U}_{g_i}\mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{g}_{f,t} &= \sigma(\mathbf{W}_{g_f}\mathbf{x}_t + \mathbf{U}_{g_f}\mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{g}_{o,t} &= \sigma(\mathbf{W}_{g_o}\mathbf{x}_t + \mathbf{U}_{g_o}\mathbf{h}_{t-1} + \mathbf{b}_o), \end{aligned} \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function. Moreover, $\{\mathbf{W}_{g_i}, \mathbf{U}_{g_i}, \mathbf{b}_i\}$ are the trained parameters for mapping from $\{\mathbf{x}_t, \mathbf{h}_{t-1}\}$ to $\mathbf{g}_{i,t}$. Similarly, $\{\mathbf{W}_{g_f}, \mathbf{U}_{g_f}, \mathbf{b}_f\}$ and $\{\mathbf{W}_{g_o}, \mathbf{U}_{g_o}, \mathbf{b}_o\}$ are the trained parameters for $\mathbf{g}_{f,t}$ and $\mathbf{g}_{o,t}$, respectively. Based on the above gates, the memory of the t-th frame $\mathbf{c}_t$ can be updated by

$$\mathbf{c}_t = \mathbf{g}_{f,t} \odot \mathbf{c}_{t-1} + \mathbf{g}_{i,t} \odot \phi(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1}), \tag{9}$$

where $\phi(\cdot)$ is the hyperbolic function, and $\odot$ stands for the component-wise product. $\mathbf{W}_c$ and $\mathbf{U}_c$ are parameters to be trained. Then, hidden states can be obtained by

$$\mathbf{h}_t = \mathbf{g}_{o,t} \odot \phi(\mathbf{c}_t). \tag{10}$$

Upon above hidden states, our LSTM outputs $\hat{\pi}_t$ and $\hat{\sigma}_t$, which approximate the targets π_t and σ_t as follows. Note that $\hat{\pi}_t = \{\hat{\pi}_t^i\}_i^5$ and $\hat{\sigma}_t = \{\hat{\sigma}_t^i\}_i^5$, corresponding to the predicted weights and standard deviations of GMs for face, left eye, right eye, nose and mouth, respectively. Since $\{\pi_t^i\}_i^5$ are continuous variables and $\sum_{i=1}^5 \pi_t^i$ is equivalent to 1, our LSTM predicts it as follows,

$$\hat{\pi}_t^i = \frac{\exp(\mathbf{W}_{\pi^i}\mathbf{h}_t)}{\sum_{j=1}^5 \exp(\mathbf{W}_{\pi^j}\mathbf{h}_t)}. \tag{11}$$

In addition, σ_t^i can be obtained by

$$\hat{\sigma}_t^i = \mathbf{W}_{\sigma^i}\mathbf{h}_t, \tag{12}$$

which is also continuous variable. In (11) and (12), $\mathbf{W}_{\pi^i}$ and $\mathbf{W}_{\sigma^i}$ are the parameters to be trained.

Besides, we define the loss function $\mathcal{L}$ of LSTM by

$$\mathcal{L} = \frac{1}{\sum_{n=1}^N T_n} \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{i=1}^5 (\underbrace{\pi_{n,t}^i \ln(\hat{\pi}_{n,t}^i)}_{\text{Cross Entropy Error}} + \underbrace{\lambda\|\sigma_{n,t}^i - \hat{\sigma}_{n,t}^i\|_2}_{\text{Mean Squared Error}}), \tag{13}$$

where N is the total number of training videos, and T_n is the number of frames for the n-th video in training set. $\pi_{n,t}^i$ and $\sigma_{n,t}^i$ are the weight and standard deviation of the GMM for the t-th frame in the n-th video, respectively. As shown in (13), the loss function includes two parts: the first part is

Table 1: Comparison of our and other approaches in mean (±standard deviation) of NSS and CC, averaged over all 7-fold cross-validation in our database.

Metrics	Our approach	Itti[19]	Cerf[6]	Judd[21]	PQFT[9]	Zhao[40]	Rudoy[31]	Xu[37]	OBDL[12]
NSS	**5.51**±1.63	1.46±0.75	2.45±0.70	1.82±0.33	1.27±0.87	4.00±1.14	1.98±0.65	4.90±1.08	1.78±1.20
CC	**0.84**±0.10	0.39±0.14	0.59±0.10	0.52±0.06	0.26±0.15	0.78±0.09	0.56±0.12	0.78±0.11	0.34±0.14

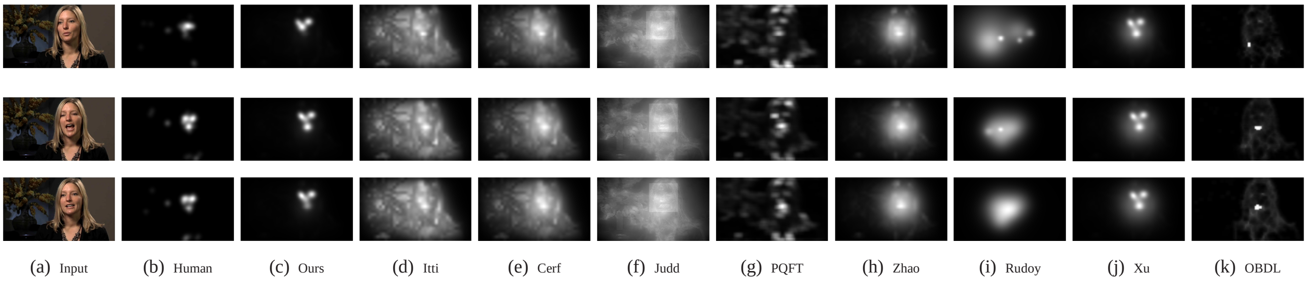

(a) Input (b) Human (c) Ours (d) Itti (e) Cerf (f) Judd (g) PQFT (h) Zhao (i) Rudoy (j) Xu (k) OBDL

Figure 9: Saliency maps across different frames (the 84th, 198th, and 203rd frames) of a randomly selected video, generated by our and other 8 approaches, as well as the human fixations.

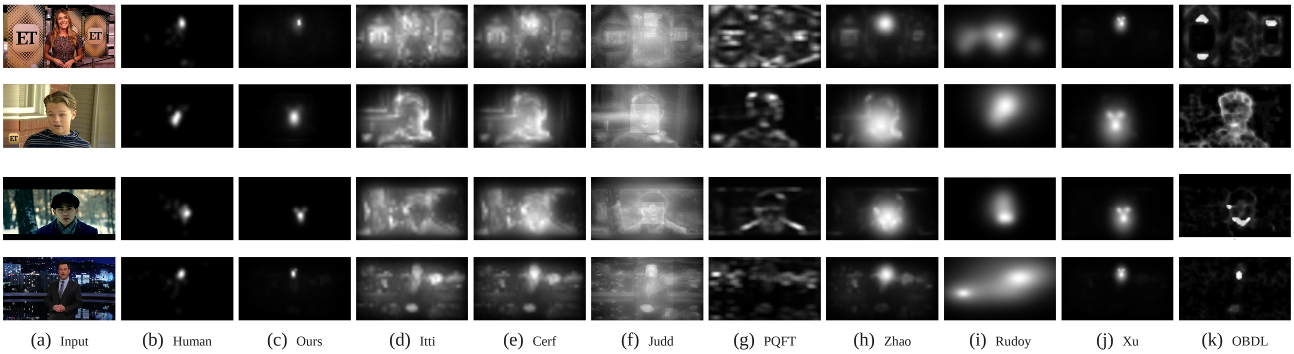

(a) Input (b) Human (c) Ours (d) Itti (e) Cerf (f) Judd (g) PQFT (h) Zhao (i) Rudoy (j) Xu (k) OBDL

Figure 10: Saliency maps of some videos at different face sizes, generated by our and other 8 approaches, as well as the human fixations.

measured by a cross entropy error for $\hat{\pi}_t$, and the second part is measured by a mean squared error for $\hat{\sigma}_t$. Besides, there is a parameter λ to regulate the weights on these two parts. Afterwards, we can simultaneously optimize $\hat{\pi}_t$ and $\hat{\sigma}_t$ using the backpropagation (BP) algorithm and train our LSTM network end-to-end. Finally, the LSTM network can be obtained for predicting DGMM distribution of visual attention within a single face across frames.

5. Experimental results

In this section, we evaluate the effectiveness of our LSTM-DGMM approach in detecting saliency of single-face videos, via comparing with other 8 state-of-the-art approaches. Next, we present the baseline and settings of our experiment.

5.1. Baseline and Settings

In our experiment, all 70 videos in our database, which are discussed in Section 2.1, were used for training and test. Here, 7-fold cross-validation was applied by randomly and equally dividing our database to be 7 non-overlapping subsets. Then, the averaged saliency detection results are reported in this section. To evaluate saliency detection results, we utilize the following metrics: NSS and linear correlation (CC). NSS quantifies the degree of correspondence between human fixation locations and saliency maps. CC measures the strength of linear correlation between human fixation maps and predicted saliency maps. The larger NSS and CC indicate higher accuracy of saliency detection.

With ground-truth fixations of each training set, π_t and σ_t of DGMM distribution can be obtained by Algorithm 1, as the targets of LSTM. In Algorithm 1, the termination criterion parameters ε and K were empirically set to 0.001 and 20, respectively. We found that Algorithm 1 converges with 6.4 iterations in average. Upon the obtained DGMM distribution, LSTM was trained to predict $\hat{\pi}_t$ and $\hat{\sigma}_t$ of DGMM. Consequently, a single layer with 30 hidden units was applied to our LSTM, in order to avoid over-fitting. The learning rate for training our LSTM was fixed to 0.003 and training epoches were set to 100, for achieving the proper convergence in training LSTM with the BP algorithm. Besides, λ, which tradeoffs $\hat{\pi}_t$ and $\hat{\sigma}_t$ in the loss function of (13), was tuned to be 0.1.

5.2. Test on our database

Objective evaluation. Here, we evaluate the saliency detection performance of 7-fold cross-validation over our database, and compare our approach with 8 other approach-

Table 2: Comparison of averaged NSS and CC (±standard deviation) for predicting fixations of other subjects.

Metrics	Our approach	Itti[19]	Cerf[6]	Judd[21]	PQFT[9]	Zhao[40]	Rudoy[31]	Xu[37]	OBDL[12]
NSS	**5.00**±1.31	1.49±1.04	2.27±1.02	1.95±0.34	1.34±1.04	2.94±1.14	2.00±0.61	4.10±0.68	1.81±1.39
CC	**0.85**±0.02	0.38±0.13	0.57±0.11	0.53±0.04	0.25±0.15	0.72±0.12	0.54±0.11	0.82±0.06	0.32±0.16

Table 3: Comparison of our and other approaches in NSS and CC, averaged over single-face videos of other database.

Metrics	Our approach	Itti[19]	Cerf[6]	Judd[21]	PQFT[9]	Zhao[40]	Rudoy[31]	Xu[37]	OBDL[12]
NSS	**3.22**	1.18	1.70	1.56	1.33	2.28	1.73	2.77	2.10
CC	**0.68**	0.37	0.48	0.35	0.29	0.55	0.41	0.58	0.46

es (i.e., Itti *et al.*[19], Cerf *et al.*[6], Juddy *et al.*[21], Guo *et al.*[9], Zhao *et al.*[40], Rudoy *et al.*[31], Xu *et al.*[37] and Hossein *et al.*[12]) to verify the effectiveness of our approach. The comparison results are presented in Table 1, in terms of averaged NSS and CC values with standard deviations. As we can see from this table, our approach outperforms other 8 approaches in terms of both two metrics, arriving at 5.51 in NSS and 0.84 in CC. Specifically, our approach has at least 0.61 improvement in NSS and 0.06 enhancement in CC, over other approaches. The results of Table 1 imply that our approach significantly advances state-of-the-art saliency detection in single-face videos. Besides, the gain of our approach over [37] verifies the effectiveness of making GMM dynamic in videos, since face saliency is modeled by GMM in [37] while it is represented by DGMM in our video approach.

Subjective evaluation. Figures 9 and 10 demonstrate the saliency maps of some selected video frames, generated by our and other 8 approaches. One may observe that the saliency maps of our approach are much closer to the ground-truth maps of human attention, compared to other 8 approaches. Such results mean that our approach is capable of well locating the salient regions within face. Specifically, Figure 9 shows the saliency maps across different frames in a same video, and we can see that our approach is able to precisely catch the saliency change when mouth is moving, while other approaches, especially [37], have nearly no reaction to this kind of movement. This clearly verifies the effectiveness of our LSTM-DGMM model, which enables the dynamic transition of GMM between frames for modeling saliency distribution within a single face. Figure 10 further shows the saliency maps of different videos with face being at various sizes. One may observe that our approach is capable of predicting human attention well, regardless of face size.

5.3. Test on generalization

Generalization on other subjects. To test the generalization of our approach, this section moves to the evaluation on predicting attention of other subjects in our database. First, we randomly selected one test set (i.e., 10 videos in total) from 7-fold cross-validation. Then, 32 subjects, totally different from those for our eye tracking database, were involved in free-viewing the selected 10 videos. Meanwhile, the fixations of those 32 subjects were recorded using the same procedure of Section 2.1. In our experiments, their fixations on 10 test videos are predicted by our LSTM-DGMM approach, which was trained from fixations of other 60 videos viewed by other 40 subjects (available in our database of Section 2.1). The accuracy of saliency detection is reported in Table 2. This table verifies the cross-subject generalization of our approach.

Generalization on other databases. We further evaluate our approach on single-face videos of other databases for generalization test. There are in total 4 videos[6] including one obvious face in the existing eye tracking database of videos, SFU [10] and DIEM [26]. They are all tested in our experiment. Note that here we randomly utilize one trained LSTM-DGMM model from the 7-fold cross-validation to test these 4 videos. Table 3 tabulates the NSS and CC results of our and other approaches, averaged over 4 test videos. From this table we can find that our approach again performs better than all other approaches. Hence, the generalization of our approach can be validated.

6. Conclusions

In this paper, we have proposed the LSTM-DGMM approach to predict saliency distribution within single-face across video frames. To be more specific, a new eye tracking database was obtained, by recording fixations of 40 subjects on viewing 70 single-face videos. To the best of our knowledge, our database is the first one for eye-tracking data of single-face videos. Then, we investigated from our database that face attracts most attention in videos, and that the distribution of attention within a face is correlated with face size, facial features and mouth movement. According to our investigation, we proposed the DGMM distribution to model the attention within a single face alongside frames. Next, benefitting from the recent development of RNN, an advanced LSTM network was developed, which predicts the structured DGMM distribution of attention on single-face videos. Finally, the experimental results showed that our LSTM-DGMM approach performs best in detecting saliency of single-face videos, compared with other 8 state-of-the-art approaches.

[6] The 4 videos are *news_tony_blair_resignation*, *ami_ib4010_closeup*, *FOREMAN* and *one_show*.

References

[1] Predicting salient face in multiple-face videos. In *Submitted to CVPR*, June 2017.

[2] J. Anderson and B. Anderson. The myth of persistence of vision revisited. *Journal of Film and Video*, 45(1):3–12, 1993.

[3] C. M. Bishop. *Pattern recognition and machine learning*. springer New York, 2006.

[4] G. Boccignone. Nonparametric bayesian attentive video analysis. In *ICPR*, pages 1–4. IEEE Computer Society Press, 2008.

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

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

[7] 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*, 52:410–432, 2016.

[8] 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*, 24(1):27–38, 2014.

[9] 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*, 19(1):185–198, 2010.

[10] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić. Eye-tracking database for a set of standard video sequences. *Image Processing, IEEE Transactions on*, 21(2):898–903, 2012.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Spt. 1997.

[12] 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*, pages 5501–5510, 2015.

[13] W. Hou, X. Gao, D. Tao, and X. Li. Visual saliency detection using information divergence. *Pattern Recognition*, 46(10):2658–2669, 2013.

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

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

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

[17] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[18] 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*, pages 64–78. International Society for Optics and Photonics, 2004.

[19] 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*, 20(11):1254–1259, 1998.

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

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

[22] J. Li, Y. Tian, T. Huang, and W. Gao. Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision*, 90(2):150–165, 2010.

[23] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen. A data-driven metric for comprehensive evaluation of saliency models. In *ICCV*, 2015.

[24] 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*, 35(2):314–328, 2013.

[25] 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*, 18(1):134–139, 2008.

[26] 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*, 3(1):5–24, Mar. 2011.

[27] 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.

[28] 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*, 17(4):564–573, 2008.

[29] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[30] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan. Regularized feature reconstruction for spatio-temporal saliency detection. *Image Processing, IEEE Transactions on*, 22(8):3120–3132, 2013.

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

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

[33] 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*, pages 50–58, 2015.

[34] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

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

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

[37] M. Xu, Y. Ren, and Z. Wang. Learning to predict saliency on face images. In *ICCV*, pages 3907–3915, 2015.

[38] 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*, pages 2944–2949. AAAI Press Cambridge, MA, 2009.

[39] 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)*, pages 491–500, 2015.

[40] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3):9, 2011.