Visual Attention on Human Face

Xiongkuo Min, Guangtao Zhai, Ke Gu

Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
{minxiongkuo,zhaiguangtao,gukesjtuee}@sjtu.edu.cn

***Abstract*—Human faces are always the focus of visual attention since faces can provide plenty of information. Although some visual attention models incorporating face cues work better in scenes containing faces, no visual attention model is particularly designed for faces. On faces, many high-level factors will influence visual attention distribution. In practice, there are many visual communication systems in which faces occupy the scenes, such as video calls. Specific visual attention model designed for face images will be of great value in these circumstances. In this paper, we conduct research on visual attention analysis and modelling on human faces. To facilitate this research, we collect 120 face images and perform eye-tracking experiments with these images. Eye-movement data shows that detailed visual attention allocation exists on faces. Using face detection and facial landmark localization, we find that some facial features are highly effective for visual attention prediction. The performance of many visual attention models can be improved by incorporating those facial features.**

***Index Terms*—Human face; visual attention; saliency map**

I. Introduction

Visual attention has been widely investigated and applied in numerous signal processing, computer vision and artificial intelligence applications. In the long term research of visual attention, various computational models have been proposed with encouraging results. Borji and Itti [1] gave a summary of state-of-the-art visual attention models. Traditional visual saliency models generally take full advantage of images' low-level features. They work well for images that can be well represented by low-level features. But they are not so efficient in some "semantic" situations, for example in some social scenes. Birmingham *et al.* [2] found that saliency did not account for fixations within social scenes, and observers' fixations were mainly driven by the social information.

Since many top-down factors are proven to be important features attracting visual attention, many researchers start to take some high-level features into consideration. Cerf *et al.* [3] incorporated face detectors into their model since they observed that faces would attract gaze independently of the task. Judd *et al.* [4] learned a saliency model based on low, middle and high-level image features. In their work, high-level features were composed of face, person and car detectors. Faces are important cues for visual attention. They can provide various types of information, such as emotion and identity. So faces are always focuses of visual attention in many scenes. Visual attention models incorporating face features generally combine low-level saliency with face detectors, and they work better in scenes containing human faces. But we find that it is efficient to emphasize "small" faces in images with simple face detection. But for images with "large" faces, more detailed visual attention allocation on faces will help, especially in close-ups of faces.

Face processing has been an active research area for years. Gaze allocation on faces is an important aspect of face processing. Some research has been devoted to understand how people gaze when viewing faces. Langton *et al.* [5] gave a review to human social attention. An important conclusion is that people tend to look at faces in scenes, especially at faces' eyes. Birmingham *et al.* [2] assessed the role of saliency in this process. They found that saliency was not the main reason accounting for the eye bias. They provided evidence that observers fixated eyes to acquire social information. Tasks also have an impact on the gaze distribution. Buchan *et al.* [6] analyzed the eye movement data collected under a speech recognition and an emotion judgment task condition. Subjects gazed more at the eyes when judging emotions, while they gazed more at the mouths when recognizing words. That was, observers were trying to achieve their goals by focusing more on particular parts of faces. Facial expression is also a factor that will influence gaze distribution. Eisenbarth and Alpers [7] found that subjects fixated mouth areas for a longer time in happy expressions, while eyes would receive more attention in sad or angry expressions. It is reasonable since some facial regions are most characteristic for specific expressions.

Despite of various visual attention models incorporating face detectors and plentiful psychology works considering gaze distribution on faces, little work has been done to build a visual attention computation models for faces. There are many practical visual communication systems in which faces occupy the scenes, such as video calls. In such systems, face-optimized visual attention models are needed. To investigate visual attention allocation on faces and build visual attention models for faces, we collect 120 images containing faces, then perform eye-tracking experiments with these images. We find that participants will fixate the "eyes", "mouths" or "noses" on large faces. It means that there are more detailed gaze distribution on large faces. Based on the eye-movement analysis, we propose to incorporate some facial features, e.g. eyes, mouth and nose. Low-level saliency and facial features are combined through learning. Comparing with state-of-the-art saliency models, effectiveness of the improved models are

IEEE VCIP'15, Dec. 13 - Dec. 16, 2015, Singapore.
978-1-4673-7314-2/15/$31.00

demonstrated based on the collected face images and eye-tracking data.

This paper is organized as follows. In Section II, we introduce the eye-tracking experiments. Gaze distribution on faces is analyzed in Section III. In Section V, low-level saliency is combined with some facial features through learning. Effectiveness of the improved models are also demonstrated in this section. Section IV concludes this paper.

II. Subjective Eye-tracking Experiments

A. Apparatus

To record the eye movement, the subjective experiments are performed with Tobii T120 Eye Tracker. Tobii T120 is a screen based eye tracker. All hardwares are integrated into a 17 inch display. The resolution of the display is 1280×1024 pixels. It can record eye movement data at a sampling rate of 60 or 120 Hz. We adopt 60 Hz in our tests. It has an effective tracking range of $50 \sim 80$ cm. Subjects are seated around 60 cm from the eye-tracker.

B. Stimuli

We collect 120 source images from Flickr. The collected images are cropped to resolutions of 1280×960, 1024×1024 or 768×1024 pixels. Collected images are mainly close-ups of faces which occupy most of the scenes. Test images contain faces of different age, gender, race, expression and etc.

C. Procedure and condition

A total of 25 subjects participate in our eye-tracking experiments. Three subjects' eye movement data are abandoned because of tracking problems, left with 22 subjects' eye-tracking data. Each test image is presented for 4 seconds, followed by a 1 second gray screen interval. Test images are showed in a random order, and the presentation order for each viewer is different. The whole test lasts around 10 minutes. All participants take part in the test in a free-viewing condition. An overview of the details of our eye-tracking experiments is listed in Table I.

TABLE I
Experiments settings and test conditions

Category	Items	Details
Eye tracker	Model	Tobii T120
	Size	17 inch
	Resolution	1280×1024 pixels
	Tracking distance	60 cm
	Sampling rate	60 Hz
Stimuli	Image number	120
	Image resolution	1280×960, 1024×1024 or 768×1024 pixels
Image presentation	Viewing time	4 seconds
	Gray interval	1 second
	Presentation order	Random
Other settings	Participants number	22 reserved 3 discarded
	Task condition	Free-viewing

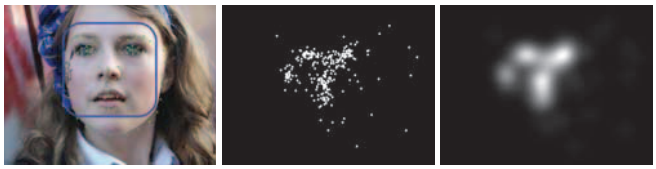
III. Gaze Distribution on Faces

A. Face detection and facial landmark localization

In this paper, we perform face detection and facial landmark localization using Face++ Research Toolkit Face++ Research Toolkit is a cloud-based SDK. Besides detecting faces and positioning facial landmarks, it can provide some face inferred information, e.g. age, gender and race. An examples of face detection and facial landmark localization is illustrated in Fig.1(a). We mainly use face position, face size and facial landmark position in this research.

B. Fixation map and fixation density map

A fixation map is generated from the eye movement data for each test image. We overlay all subjects' fixations in a binary map. In this map, the fixated positions are set to 1 while other positions are set to 0. The fixation maps are then filtered with a Gaussian kernel, generating the fixation density maps (FDMs). The standard deviation of the Gaussian kernel is set to around 1 degree of visual angle, i.e. $\sigma = 40$. Fig.1(b) and Fig.1(c) give an illustration of fixation map and corresponding fixation density map.



(a) (b) (c)

Fig. 1. (a) Face detection and facial landmark localization, (b) fixation map, (c) fixation density map.

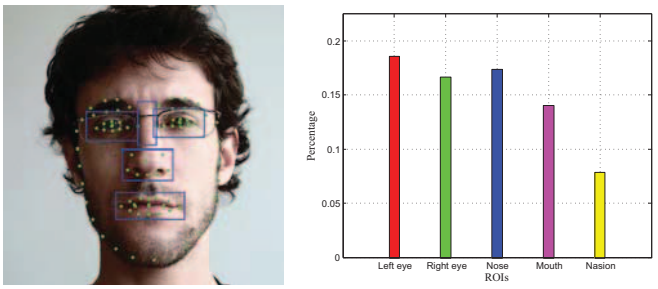
C. Overall gaze distribution

Since we are interested in gaze distribution on faces, we create several rectangular regions of interest (ROIs) for statistical analysis: left eye, right eye, nose, mouth and nasion regions. The nasion area is among the eyes and nose. Certain amounts of fixations fall on this area, so we take this area as a ROI too. Fig. 2(a) illustrates an example of ROIs. Areas within the rectangless are created ROIs. Green points represent facial landmarks. The locations and sizes of rectangular ROIs are set to cover the key facial landmarks. We analyze the gaze distribution on faces based on collected eye-tracking data. Fig. 2(b) illustrates the overall results. The vertical axis represents the percentage of fixations fixated corresponding ROI. When fixation fall on the overlapping ROIs, it is treated as fixation belonging to both ROIs. From Fig. 2(b), we can see that the regions of eyes and nose are the most salient parts of faces.

IV. Combining low-level saliency with facial features

A. Facial features

As described in Section III-C, fixations mainly fall on several particular areas such as eyes and noses. Traditional low-level saliency can not predict such kind of high-level



(a) (b)

Fig. 2. (a) Example of ROIs. Areas within the ellipses are created ROIs. The green points are facial landmarks. (b) Overall gaze distribution on faces.

semantic information well. So we propose to combine low-level saliency with several facial features. In this section, the high-level facial features used in this paper will be introduced.

We introduce 5 facial features corresponding to 5 ROIs analysed in Section III-C. Fig.3 shows an example of facial features for a test image. Three images are illustrated for each ROI. The first and second images show cropped ROI and FDM. Certain degrees of center bias can be revealed from FDMs. The bottom right of left eye region, bottom left of right eye region, middle of nose and mouth regions tend to be more salient. The facial feature maps are calculated by simply placing uniform Gaussian kernels at the positions of several facial landmarks. The standard deviation of the kernel is set to 40 (around 1 degree of visual angle). The chosen landmarks are marked as red in Fig.3. To match the center bias described above, the landmarks for each feature are chosen mainly according to its spatial locations. For the eyes, 4 landmarks in the corner closer to the face center are chosen. For nasion, three points distributed along the nasion and nose bridge are calculated from existing landmarks. For nose and mouth, 2 and 4 landmarks in the center are chosen respectively. The third image for each ROI shows the computed facial feature map. The introduced facial features provide fine correlations to FDMs.

B. Combining saliency with facial features through learning

We adopt a learning approach similar to [4] to combine traditional saliency with facial features. We trained a classifier from the eye-tracking data using a 10-fold cross validation. We have 108 training images and 12 test images in each validation. For each training image, we randomly select 10 samples from top 20% and bottom 70% salient areas respectively, generating a training set of 1080 positive and 1080 negative samples. We consider two kinds of features in this paper. One kind is saliency computed from traditional models [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [3], [4]. Among these models, calculated saliency is low-level and bottom-up saliency except for SMVJ [3] and Judd *et al.* [4]. SMVJ [3] and Judd *et al.* [4] contain high-level object detectors. Another kind is high-level facial features described in Section IV-A. At last, a total of 6 features (1 traditional low-level saliency and 5 facial features) are used in the learning process. Feature values at the positions of chosen samples are concatenated into a feature vector. The

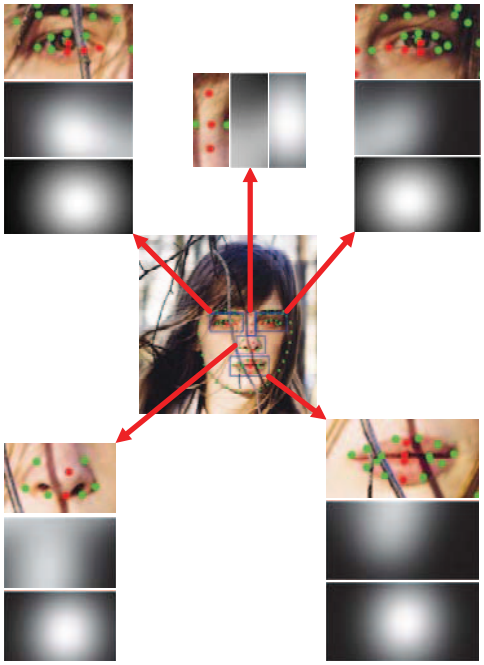


Fig. 3. An example of facial features. Three images are illustrated for each ROI. From top/left to down/right: cropped ROI, FDM and calculated facial feature map. Green points: facial landmarks; red points: points used to calculated facial features.

feature vectors are then used to train a model through liblinear support vector machine [18]. The sample choosing and model training approaches are similar to [4]. We only replace the features and adjust some parameters accordingly.

C. Experimental results

We mainly compare the performance of traditional saliency models and the improved models which integrate traditional saliency and facial features. We conduct learning and testing experiments on VAF database. The performance of all models are evaluated by sAUC, CC and NSS. Table II lists the experiment results. Traditional saliency models considering some higher-level features generally perform better. For example, SMVJ [3] and Judd [4] both used face detectors. Moreover, when combined with facial features, all models perform better. The performances of all models are promoted to a certain degree since facial features are highly effective in face images. In addition to traditional models and corresponding improved models, we also implement several other simple but effective models. As listed in Table II, "GaussIC" and "GaussFC" denote the Gaussian models whose Gaussian blob is located in the image and face center respectively. "Face" denotes the model that only combines 5 facial features through learning. "Human" denotes the human model which use saliency maps calculated from other 21 subjects to predict fixations of 1 subject under test. We do not list the CC scores of the human model because the correlation between FDMs of one subject and 21 subjects is not good. GaussFC performs better than GaussIC since the center of gaze is always the face center. But the face center will deviate from the image center. Performance

TABLE II
PERFORMANCE OF TRADITIONAL (TRAD) AND CORRESPONDING IMPROVED (IMPR) SALIENCY MODELS.

Models	sAUC		CC		NSS	
	TRAD	IMPR	TRAD	IMPR	TRAD	IMPR
AIM [8]	0.6178	0.7241	0.2719	0.8472	0.7962	2.6837
AWS [9]	0.6377	0.7231	0.3113	0.8466	0.9596	2.6837
CA [10]	0.6071	0.7226	0.2374	0.8497	0.6907	2.6855
Hou [11]	0.5771	0.7204	0.4771	0.8517	0.6963	2.6994
IT [12]	0.5942	0.7145	0.4239	0.8183	1.1446	2.5151
SeR [13]	0.6071	0.7201	0.1908	0.8540	0.5891	2.7288
SR [14]	0.5987	0.7197	0.1953	0.8513	0.5604	2.7158
SUN [15]	0.6180	0.7216	0.2274	0.8525	0.6769	2.7256
Torra. [16]	0.6080	0.7235	0.2202	0.8429	0.7007	2.6887
GBVS [17]	0.6077	0.7132	0.4771	0.7806	1.3079	2.3348
SMVJ [3]	0.6537	0.7079	0.5870	0.7467	1.5867	2.1504
Judd [4]	0.6226	0.7007	0.5885	0.7308	1.5367	2.0578
GaussIC		0.5841		0.7811		2.0943
GaussFC		0.6603		0.8452		2.2752
Face		0.7194		0.8561		2.7525
Human		0.7349		-		2.9923

of the Face model is comparable to state-of-the-art models listed in Table II. It indicates that what account the most for visual attention distribution on faces are those high-level facial information. Fig. 4 illustrates an example of sample image, FDM, and saliency maps calculated from 4 chosen models listed in Table II. During experiments, we also find that saliency maps of improved models look similar. So low-level saliency only contributes little to the final saliency maps.

V. CONCLUSION

In this paper, we investigate the problem of visual attention distribution on human faces. We perform eye-tracking experiments on faces of various sizes. We collect 120 face images and perform eye-tracking experiments with these images. Overall fixation distribution on faces is analyzed. Subjects generally fixate on particular areas, e.g. eyes, mouth and nose. We evaluate state-of-the-art saliency models with collected face images and corresponding eye-movement data. Not surprisingly, models incorporating face cues perform better. Through face detection and facial landmark localization, we introduce several high-level facial features. When combined with facial features, visual attention models can predict human fixations much better on faces. So in the study of visual attention for face images, more high-level factors should be taken into consideration.

ACKNOWLEDGMENT

This work was supported in part by NSFC (61025005, 61371146, 61221001), 973 Program (2010CB731401) and FANEDD (201339).

REFERENCES

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

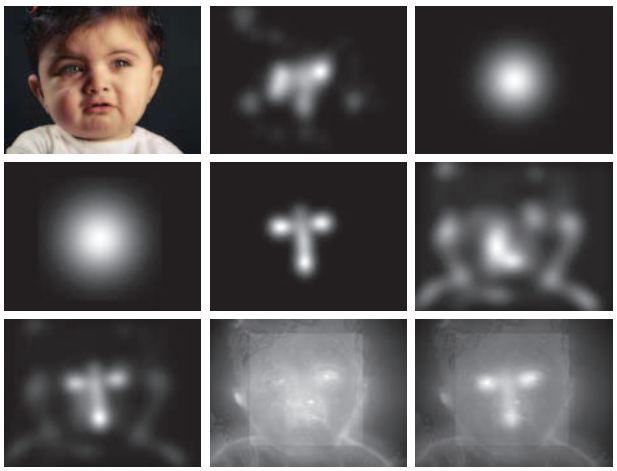


Fig. 4. Saliency maps of sample image. From left to right, from top to bottom: sample image, FDM, GaussIC, GaussFC, Face model, IT [12], Improved IT, Judd [4], Improved Judd.

[2] E. Birmingham, W. F. Bischof, and A. Kingstone, "Saliency does not account for fixations to eyes within social scenes," *Vision Research*, vol. 49, no. 24, pp. 2992–3000, 2009.
[3] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*, 2008, pp. 241–248.
[4] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 2106–2113.
[5] S. R. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50–59, 2000.
[6] J. N. Buchan, M. Paré, and K. G. Munhall, "Spatial statistics of gaze fixations during dynamic face processing," *Social Neuroscience*, vol. 2, no. 1, pp. 1–13, 2007.
[7] H. Eisenbarth and G. W. Alpers, "Happy mouth and sad eyes: scanning emotional facial expressions." *Emotion*, vol. 11, no. 4, p. 860, 2011.
[8] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2005, pp. 155–162.
[9] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.
[10] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
[11] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2009, pp. 681–688.
[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
[13] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, 2009.
[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.
[15] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.
[16] A. Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.
[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.
[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.