Estimate Gaze Density by Incorporating Emotion

Huiying Liu[1], Min Xu[2], Xiangjian He[2], and Jinqiao Wang[3]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]University of Technology, Sydney, Australia
[3]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
liuhy@i2r.a-star.edu.sg, min.xu25@gmail.com, xiangjian.he@uts.edu.au, jqwang@nlpr.ia.ac.cn

ABSTRACT

Gaze density estimation has attracted many research efforts in the past years. The factors considered in the existing methods include low level feature saliency, spatial position, and objects. Emotion, as an important factor driving attention, has not been taken into account. In this paper, we are the first to estimate gaze density through incorporating emotion. To estimate the emotion intensity of each position in an image, we consider three aspects, generic emotional content, facial expression intensity, and emotional objects. Generic emotional content is estimated by using Multiple instance learning, which is employed to train an emotion detector from weakly labeled images. Facial expression intensity is estimated by using a ranking method. Emotional objects are detected, by taking blood/injury and worm/snake as examples. Finally, emotion intensity, low level feature saliency, and spatial position, are fused, through a linear support vector machine, to estimate gaze density. The performance is tested on public eye tracking dataset. Experimental results indicate that incorporating emotion does improve the performance of gaze density estimation.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms

Keywords

Visual saliency; visual attention; emotion

1. INTRODUCTION

In the past years, gaze density estimation, as an aspect of visual attention analysis, has attracted many research efforts for its wide applications. Attention is driven by both stimulus and task. Stimulus driven attention is also called bottom-up attention. The representative works include Itti's saliency based method [5], Wang's Site Entropy Rate method [14], attention based on Information Maximization (AIM) method [1], the Saliency Using Natural (SUN) images method [15]. Task driven attention is also known as top-down attention. The representative works include [6] and others. In most of the existing works, the ”task” usually refers to object detection, object recognition, scene analysis, and so on. Playing a significant role, emotion drives attention (examples are shown in Fig. 1) but has not been considered for gaze density estimation.

This paper is the first attempt to explore the influence of emotion on gaze. There are many psychology studies that suggest emotion plays an important role in guiding attention [9][12]. The major difficulty of this work is to estimate the emotion intensity of each position of an image. Our solution proposes a learning based emotion detector as major component. The detector is trained through Multiple Instance Learning (MIL), from a weakly labeled dataset, i.e., the images are labeled with emotion type but not the boundary of the emotional content. In addition to the emotion detector, we adopt two types of specific detectors/estimators as

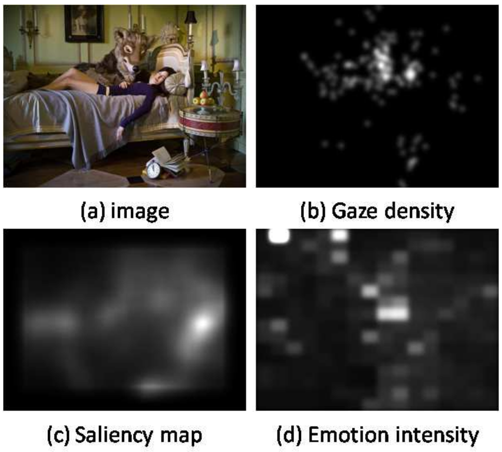


Figure 1: An examples of attention driven by emotion. As shown in the image (a) and the saliency map (c)[14], from the point of view of low level features, the beast is not the most salient part. But as shown in the gaze density map (b), it does attract much attention. The reason is that it evokes emotion, emotion of fear. The emotion intensity is shown in (d).

MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654983.

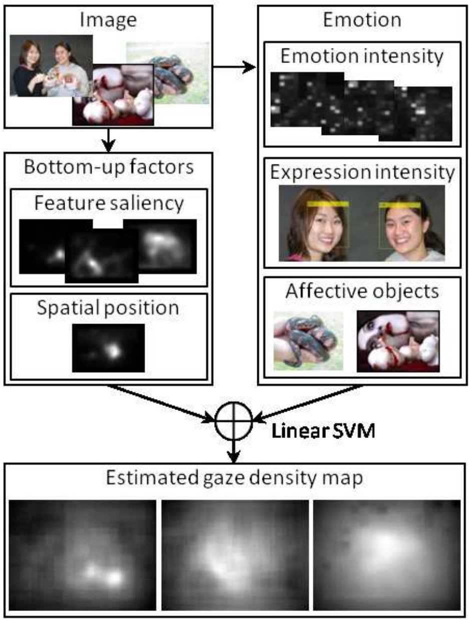


Figure 2: The overview of the method.

complement. One is an expression intensity estimator and the other contains two object detectors, e.g., worm/snake detector and injury detector. After obtaining the emotion intensity, linear Support Vector Machine is applied to fuse emotion with low level feature saliency and spatial position for gaze density estimation.

The major contribution of this paper is that, emotion factor is for the first time considered, as a complementation with the traditional factors, for gaze density estimation. Another contribution is that we propose an emotion intensity estimation method, by using multiple instance learning. As far as we know, this is also the first attempt. Experiments performed on images with rich emotions show that the incorporation of emotion does improve the performance of the baseline methods.

2. INCORPORATING EMOTION INTO GAZE DENSITY ESTIMATION

In this paper, three factors, including spatial position, feature saliency, and emotion are considered for gaze density estimation. For the factor of spatial position, the distance from a pixel to the image center is used as feature. For feature saliency, we adopt the existing methods as baseline, including Itti's method [5], AIM [1], SUN [15], and the SER method [14]. After calculating all the three factors, we fuse them through linear support vector machine to estimate the final gaze density map. Fig. 2 illustrates the overview of our method. In the following sections of the paper, we will focus on the emotion part since this is the most important contribution of our work.

We consider two types of affective images. One type includes the ones representing emotion, e.g., facial expression and body gesture. The other one consists of the images evoking emotion. For example, images of flowers usually evoke the emotion of joyfulness and the ones of snakes often evoke fear. To evaluate the emotion intensity of each position of an image, we consider three aspects. 1) Multiple instance learning is employed to train generic emotion detectors from emotional images. 2) Rank learning is adopted to estimate the intensity of facial expression. 3) We detect several emotional objects, e.g., injury, worm/snake. These three aspects will be detailed in the following subsections.

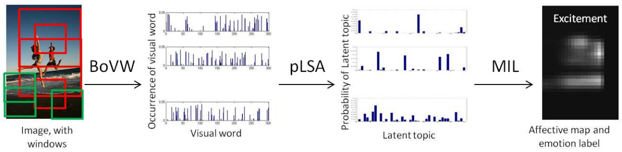

BoVW
pLSA
MIL
Excitement
Image, with windows
Occurrence of visual word
Visual word
Probability of Latent topic
Latent topic
Affective map and emotion label

Figure 3: The flowchart of the emotion intensity estimation method.

2.1 Emotion intensity estimation using MIL

To estimate the emotion intensity of each position of an image, we face two major difficulties. The first one is that emotion is so complex that an emotion can be represented by diverse contents. For instance, an image of a beach with sunshine evokes happiness, blooming flowers evokes happiness, a smiling face represents happiness, and jumping is also able to represent happiness. The second difficulty is that it is difficult to collect a training dataset with the emotional content of each image exactly marked out. For the first difficulty, we appeal to probabilistic Latent Semantic Analysis (pLSA) [4] to learn a set of latent topics from affective images. For the second one, we employ multiple instance learning [8] to weaken the need for exact labeling. The profit of using MIL is that it doesn't need exact region label but can learn from weak labels.

The problem of emotion intensity estimation is defined as: given a set of training images, each of them labeled with an emotion type, such as happy, sad, and angry, our target is to train a set of emotion detectors that can be used to estimate the emotion intensity of each position of an arbitrary image. The flowchart of the proposed method is shown in Fig. 3. For a training image, image windows are extracted at multiple scales and multiple positions thus at least one of them covers the emotional content. All the windows extracted from an image compose a bag. The Bag-of-Visual-Words (BoVW) method [3] is adopted to represent image windows. Then probabilistic Latent Semantic Analysis is adopted to represent a window with latent topics. Finally, MIL is used to train from the data an emotion detector. For a given image, we estimate the emotion state of each window using the trained model.

2.1.1 BoVW description and pLSA representation

The emotional state of an image is determined by both the global atmosphere and local contents. In our work, our purpose is to estimate the emotion intensity of each position of an image, e.g., the local content. Therefore local features are more suitable. The feature used by [7], which are delicately designed from the perspective of art and psychology, are mostly global features suitable for classification rather than for intensity estimation. SIFT descriptor is applied as basic feature in our work. Bag-of-Visual-Words is adopted to represent the image windows. Let $W = \{w_1, w_2, ..., w_M\}$ be a vocabulary, with M as the total number of visual words, each image window is represented as $\mathbf{x} = [x_1, x_2, ..., x_M]'$, with x_i indicating the occurrence of word w_i.

For each image window, we further extract latent topics using pLSA, which is verified to be effective for image scene classification [3]. pLSA model associates an unobserved class variable $\mathbf{z} = (z_1, z_2, ..., z_K)$ with each observation (I_i, w_j). K is the number of topics. With this latent variable, the probability of an observation pair (I_i, w_j), which is the occurrence of the visual word w_j in a particular image window I_i, is

$$P\left(I_i, w_j\right) = P\left(I_i\right) P\left(w_j|I_i\right) = P\left(I_i\right) \sum_{k=1}^{K} P\left(w_j|z_k\right) P\left(z_k|I_i\right) \tag{1}$$

$P(w_j|z_k)$ and $P(z_k|I_i)$ are estimated by using Expectation-Maximization (EM) algorithm.

Given an sample, the log probability of a topic is

$$\log P(z_k|I_i) = \sum_{j=1}^{M} \frac{x_{ij}}{\|\mathbf{x}_i\|_1} P\left(z_k|I_i, w_j\right) \tag{2}$$

here $P(z_k|I_i, w_j)$ is calculated as

$$P\left(z_k|I_i, w_j\right) = \frac{P\left(w_j|z_k\right) P\left(z_k|I_i\right)}{\sum_{k=1}^{K} P\left(w_j|z_k\right) P\left(z_k|I_i\right)} \tag{3}$$

Using Eq. (2) and Eq. (3), we estimate the topic distribution of each image window.

2.1.2 *MIL estimation*

In many cases, the emotion conveyed by an image is mainly contained in an local patch instead of the whole image. For example, for the leftmost image in Fig. 3, it is far-fetched to say that the patches with green borderlines are exciting ones. But the available labeling information is only for the whole image not for each image window. Multiple instance learning is a powerful method to learn from this kind of weakly labeled data[8]. For MIL, each bag of samples can share one label. If there is at least one positive sample in a bag, then the whole bag is labeled as positive. If a bag is labeled as negative, all the samples in the bag are negative.

We adopt an iterative method to train the detectors. At the initializing stage, the image window at the original level is adopted as positive sample to train the detectors. At the following step, the probability of each window to cover the emotional region is estimated, by using the trained detector. The one with the highest probability in a bag is taken as positive to train a new detector. This process is repeated until no change happens when choosing the positive samples.

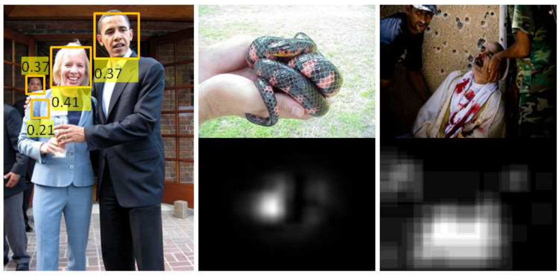
By using the trained detectors, each image window obtained in section 2.1.1 is assigned a probability of belonging to a particular emotion category. Using the result, we calculate the probability of a pixel belonging to an emotion category. For each pixel in an image, the probability of belonging to a particular emotion category is the average of all the windows covering it.

$$P(i, j, c) = \sum_{(i,j)\in\mathbf{x}_k} P(\mathbf{x}_k, c) \tag{4}$$

here (i, j) is a pixel and c is an emotion category, $\mathbf{x}_k$ is a window.

The probability of a pixel to belong to the emotional region is the summation over all the emotion categories.

$$P(i, j) = \sum_{c} \sum_{(i,j)\in\mathbf{x}_k} P(\mathbf{x}_k, c) P(c) \tag{5}$$



(a) Expression intensity (b) Snake detection (c) Injury detection

Figure 4: Examples of relative expression and emotional object detection.

here $P(c) = \max_k P(\mathbf{x}_k, c)$ is the probability of the image to belong to category c. Fig. 1 (d) shows an example of emotion intensity estimation result.

2.2 Facial expression intensity estimation

Human faces attract much human attention. This factor has been considered in many works[6] but what is not considered is that facial expression attracts more attention than the faces without expression. Generally speaking, the more intense the expression is, the more attention it attracts. This factor is particularly important for the images with multiple faces.

We borrow the idea of relative attribute to estimate expression intensity [11]. GIST is adopted as feature and rank SVM is employed to train the estimator[10]. The expressions considered are the six ones thought to be recognized universally. They are happy, anger, disgust, fear, surprise, and sad [2]. Fig. 4 (a) shows an example of relative expression. For each detected face, expression intensity is estimated for the six kinds of expression. Then the max of them is applied for gaze density estimation.

2.3 Other emotional object detection

Many objects can evoke emotion. For instance, flowers usually evoke the emotion of happiness, worms evoke disgust and fear, snakes evoke fear, etc. These objects are referred to as emotional objects in this paper. It is hardly to collect a full list of emotional objects so we choose two examples as representative. As shown in [9], fear drives attention more than other emotions. Therefore we choose two objects evoking fear as examples. They are worm/snake and blood/injury. Worm and snake look alike and they evoke the similar emotion, fear. Therefore we train one detector for them. We adopt GIST as feature and employ SVM to train a worm/snake and an blood/injury detector [10]. A sliding window is used to detect the target objects in the images. Finally, we obtain the probability of each pixel to belong to the object. Fig. 4 (b) and (c) show examples of snake detection and blood detection results.

3. EXPERIMENTS

We test our method on the NUSEF eye gaze dataset [13] for the affluent emotion contained in the images. The 446 images publicly available are used in experiment. The emotion intensity estimator are trained on the dataset published in [7].

The size of the vocabulary of BoVW is set as 100, the number of latent topic of pLSA is set as 20. For the lin-

Table 1: Test result of the components.

	ITTI	+MIL	+EXP	+FACE	+EMO
AUC	0.5744	0.5984	0.5893	0.5717	0.7498
t-test		<0.01	<0.01	<0.01*	<0.01

*The t-test result is between +FACE and +EXP.

Table 2: The result of comparison between our method and the baseline methods.

	ITTI	AIM	SUN	SER
baseline	0.5744	0.6727	0.6159	0.7425
with emotion	0.7122	0.7163	0.7020	0.7541
t-test	<0.01	<0.01	<0.01	0.04

ear SVM to fuse all the components, we choose from each training image 20 positive samples randomly from the top 5 percent, and 20 negative samples randomly from the bottom 30 percent. Cross validation is adopted with 50 samples are chosen as test data and the others (about 400 ones) are left for training data.

3.1 Test of the separate components

We first test the effectiveness of the separate components, the MIL based emotion intensity method and the relative expression intensity. For this test, the classic saliency method [5] is adopted as baseline. In our method, we evaluate relative expression intensity for gaze density estimation. Then the question is, is it better than using face directly, since there are many methods use face as a feature for gaze density estimation? To answer this question, we compare the performance of using face and the one using expression. Five sets of results are compared. 1) The baseline saliency method (ITTI). 2) The baseline method with the emotion intensity estimated using MIL (ITTI+MIL). 3) The baseline method with the expression intensity (ITTI+EXP). 4) The baseline method with human face. This is to test if the expression intensity have contributions compared with face (ITTI+FACE). 5) The baseline method with all the components, i.e., the spatial position, MIL emotion intensity, expression intensity, and emotional object detection (ITTI+EMO). The area under the ROC curve and the t-test results are shown in Table 1. The results show that the components are effective and the t-test results show the significance of the improvement.

3.2 Test of the performance

The major contribution of this paper is to incorporate emotion into gaze density estimation. Thus we test the performance of our work on the basis of existing saliency methods. The methods chosen as baseline are 1) Itti's method[5], 2) the AIM method [1], 3) the SUN method SUN [15], and 4) the SER method [14]. Table 2 shows the area under the curves and the t-test result. From the results, we can see that emotion does improve the performance of the saliency methods significantly. Fig. 5 shows two examples. We can see that our method well compliments with saliency and emphasizes the emotional content.

4. CONCLUSION

In this paper, we are the first to incorporate emotion into attention analysis, especially gaze density estimation. Experimental results on eye gaze dataset containing emotion rich images demonstrate the effectiveness of the proposed method. This is the very first attempt to incorporate emotion into attention analysis. Experiments verified the effectiveness of the proposed method.

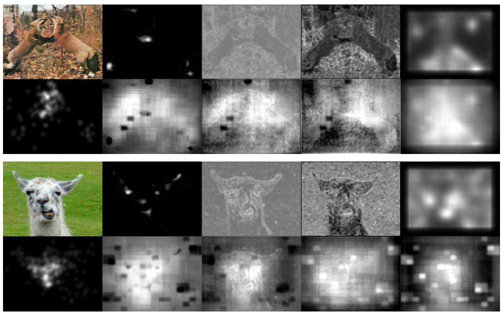


Figure 5: Two examples of gaze density estimation results, compared with the baseline methods. The first columns shows the image (top) and the user gaze density (bottom). The other four columns show the baseline methods (top) and our results (bottom). From left to right, the baseline method are itti's method [5], AIM [1], SUN [15], and SER [14].

5. ACKNOWLEDGMENT

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61273034, 61332016).

6. REFERENCES

[1] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006.

[2] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[3] L. Fei-Fei and P. Perona. *A Bayesian hierarchical model for learning natural scene categories*. CVPR, 2005.

[4] T. Hofmann. *Unsupervised learning by probabilistic latent semantic analysis*. Machine Learning, vol. 42, no. 1-2, pp. 177-196, 2001.

[5] L. Itti, C. Koch, and E. Niebur. *A model of saliency-based visual attention for rapid scene analysis*. T-PAMI, vol. 20, no. 11. pp. 1254-1259, 1998.

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

[7] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM Multimedia*, pages 83–92, 2010.

[8] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning*, pages 341–349, 1998.

[9] A. Ohman, A. Flykt, and F. Esteves. Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychoogy: General*, 130(3):466–478, 2001.

[10] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–29, 2006.

[11] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510, 2011.

[12] H. T. Schupp, J. Stockburger, M. Codispoti, M. Junghofer, A. I. Weike, and A. O. Hamm. Selective visual attention to emotion. *Journal of neuroscience*, 27(5):1082–1089, 2007.

[13] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. *An eye fixation database for saliency detection in images*. ECCV, 2010.

[14] W. Wang, Y. Wang, Q. Huang, , and W. Gao. Measuring visual saliency by site entropy rate. In *CVPR*, 2010.

[15] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, , and G. W. Cottrell. *SUN: A Bayesian framework for salience using natural statistics*. Journal of Vision, 8(7):32, 1-20, 2008.