LEARNING IMAGE SALIENCY FROM HUMAN TOUCH BEHAVIORS

Shaomin Fang[1], *Yijuan Lu*[1], *Xinmei Tian*[2]

[1]Department of Computer Science, Texas State University-San Marcos, {s_f5, lu}@txstate.edu
[2]University of Science and Technology of China, xinmei@ustc.edu.cn

ABSTRACT

The concept of touch saliency was recently introduced to generate image saliency maps based on human simple zoom behavior on touch devices. However, when browsing images on touch screen, users tend to apply a variety of touch behaviors such as pinch zoom, tap, double tap zoom, and scroll. Do these different behaviors correspond to different human attentions? Which behaviors are highly correlated with human eye fixation? How to learn a good image saliency map from various/multiple human behaviors? In this work, we design and conduct a series of studies to address these open questions. We also propose a novel touch saliency learning approach to derive a good image saliency map from a variety of human touch behaviors by using machine learning algorithm. The experimental results demonstrate the validity of our study and the potential and effectiveness of the proposed approach.

Index Terms— Touch saliency, visual saliency, touch behaviors.

1. INTRODUCTION

Visual attention refers to selective concentration on meaningful region of a scene [1]. A visual saliency map displays the spotlights of the concentrations. Visual attention learning is widely applied in image compression [2], image segmentation [3], image retargeting [4], and information retrieval [1].

In the traditional visual attention study, users' eye fixation data are required and the eye tracking device is the only equipment to collect the data. Although eye tracking has been developed for years, it is not widely popularized due to three major reasons: 1) high cost; 2) complicated operation, which requires non-trivial calibration, validation, and chin-and-forehead-rest for stabilization; 3) low mobility (not easy to carry it everywhere due to considerable size and weight).

Recently, with the popularity of touch screen phones, tablets and laptops, more and more people rely on them for daily image or video browsing, sharing, and surfing. When using a limited size of touch screen for image browsing, users tend to tap, pinch zoom, double tap zoom, and scroll to have a closer view of a particular region of interest. These touch behaviors may indicate user interests on certain regions of a image, and perhaps capture similar information as eye fixations in visual attention study. A recent study [5] called "Touch Saliency" investigated generating saliency maps based solely on simple zoom behavior. Many interesting questions naturally arise: 1) Do different touch behaviors (tap, pinch zoom, double tap zoom, scroll *etc*.) correspond to different human attentions? 2) Which behaviors are more correlated with human eye fixation? 3) How to learn a good image saliency map from various/multiple human touch behaviors?

To address these questions, we design and conduct a series of studies with the conventional eye-fixation based saliency served as ground truth. An image browsing app is designed on a touch mobile phone to collect users' touch behavior data. A novel touch saliency learning approach is also proposed to derive a good image saliency map from a variety of human touch behaviors. During the process of building a supervised learning model, the weights of different human touch behaviors are learned, which indicate the different contributions of these behaviors to the image saliency information. The experimental results demonstrate the validity of our study and the potential and effectiveness of the proposed approach.

Compared with eye-tracking devices, touch devices are much more popular, cheaper and also easier to operate and carry. The users finger behaviors are much easier to be recorded than eye-movements. Therefore, touch saliency can be easily obtained and it will definitely have wide applications in image compression, image segmentation, and image retargeting *etc*. in the near future.

The main contributions introduced in this paper are summarized as follows: 1) We quantitatively study and analyze human attention from a variety of touch behaviors in this paper, and then propose a set of valuable features from the touch information. 2) We propose to utilize a supervised learning method to automatically learn the correlation between different touch behaviors and human eye fixations, and then to derive a good image saliency map from a variety of touch behaviors. 3) Our work will guide the research in touch saliency ability estimation and opens broad research possibility for touch-based visual attention learning.

2. RELATED WORK

Xie et al. [6] made the first attempt to extract user attention by analyzing the touch information on images in 2005. They

collected data from 10 subjects on 26 images. Several attributes are considered in the users attention learning including region of interest, minimal allowable spatial area of the attention, minimal duration of the attention *etc*. This study demonstrates that users attention can be easily obtained from touch behaviors. However, its performance is not quantitatively evaluated. Therefore, its validity is unknown.

In 2012, *Xu et al.* [5] introduced a new concept of touch saliency, which is to generate image saliency maps based on a human simple zoom behavior. In their data collection, 16 participants freely viewed 440 images in NUSEF database [3] on a touch-screen mobile device. The center point of the screen is treated as the fixation point and the zoom scale is used as Gaussian filter parameters to generate the touch saliency map. This study shows that touch saliency map and visual saliency map are highly correlated with each other in an image browsing task. However, in this work, the image pixel of center point of the screen is selected as the fixation point, which always causes some bias in the saliency map learning. It is observed that when the image is zoomed in, the users do not always adjust the most salient area to the center of the screen.

In our preliminary study, we observe that when browsing images on the limited size touch screen, users tend to apply a variety of touch behaviors, such as tap, pinch zoom, double tap zoom, and scroll to find a particular region of interest and look them closer. What correlations between these different behaviors and human attention are, whether they contribute equally to the human eye fixation, and how to learn good image saliency maps from multiple touch behaviors have not been explored in the existing works. To our best knowledge, this is the first attempt that conducts a series of studies to explore these questions.

3. IMAGE SALIENCY LEARNING FROM TOUCH BEHAVIORS

3.1. Touch Behaviors Data Collection

In order to collect user touch behavior data, we develop an image browsing interface on a multi-touch mobile phone. The interface is designed as same as most popular image browsers which support tap, pinch zoom, double tap zoom, scroll, *etc*.

The same data set NUSEF [3] used in the work [5] is chosen in our study by considering its two unique attributes. First, this data set contains 446 images (size is around 1024x768 pixels) and corresponding ground truth eye fixation data acquired from an eye-tracking device with a pool of 75 subjects. Second, the images in this dataset are manually collected from Flickr, Photo.net, Google Images and IAPS, and they are representative of various semantic concepts, scales, orientations and illuminations [3].

15 users (4 females, 11 males) with the age between 24 and 33 ($\mu = 26.6$, $\sigma = 2.75$) participated our user study. Each participant freely viewed all the 446 images on the Samsung Galaxy S3 Android phone (4.8 inch HD Super AMOLED display with 1280x720 pixels). Each user can use any touch behavior to move to a particular region of interest. During the image browsing process, each image is displayed for 12 seconds, a black screen is shown for 2 seconds between any two consecutive images to avoid interference. For every user, all images are displayed in a random order. Thus, the display orders may be different for each participant to avoid bias. The program keeps recording the touch behavior type, center pixel coordinates of the pinch zoom, double tap coordinates, pixel coordinates of center point of the screen, scroll target position, tap point coordinates, and *etc*.

3.2. Touch Behaviors Features

In order to learn the relationship between different touch behaviors and the human attention on the images, we analyze all the touch behaviors data collected from the user study and propose the following five features that may indicate humans interest and attention on certain regions of the image.

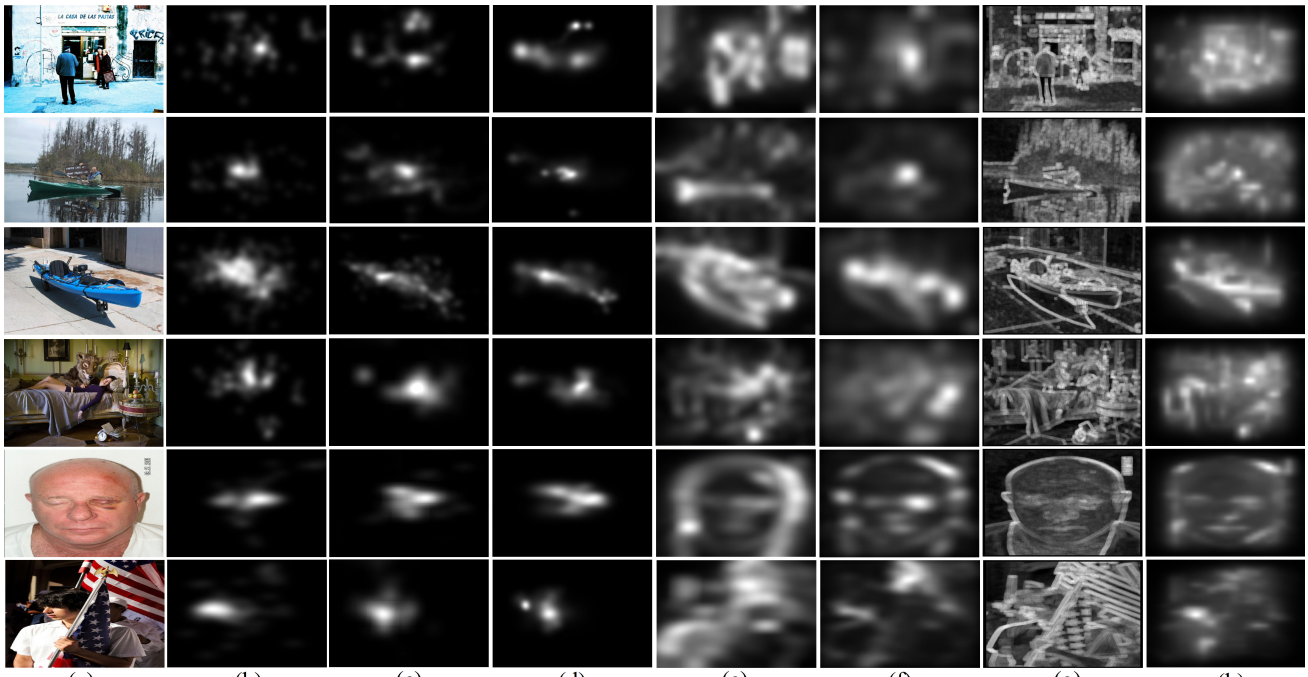
- Tap (T): Image pixel coordinates of the tap point.
- Pinch-zoom-in (P): Image pixel coordinates of the center point between two fingers after zoom in.
- Scroll (S): Image pixel coordinates of the scrolling point after zoom in.
- Double-tap-zoom-in (D): Image pixel coordinates of the double-tap point and the zoom scales of the double-tap zoom in/out on images.
- Center (C): Image pixel coordinates of the center point of the touch screen after zoom in.

3.3. Touch Saliency Learning

Different from previous touch saliency generation methods, a novel learning based approach is proposed to generate image saliency maps from the touch behaviors data.

Let $R=\{I_1,I_2,I_3,...,I_m\}$ be a set of training images. We divided an image I_k into a by b grids, $G_{I_k} = \{g^1_{I_k}, g^2_{I_k}, g^3_{I_k}, ..., g^{ab}_{I_k}\}$, where $g^j_{I_k} = (g^{T_j}_{I_k}, g^{P_j}_{I_k}, g^{S_j}_{I_k}, g^{D_j}_{I_k}, g^{C_j}_{I_k}) \in R^5$ is a touch feature vector extracted from the j-th grid. The value of these five touch behavior features $g^{T_j}_{I_k}, g^{P_j}_{I_k}, g^{S_j}_{I_k}, g^{D_j}_{I_k}, g^{C_j}_{I_k}$ are calculated by counting the number of occurrences the corresponding behavior happens in the j-th grid of image I_k. For example, if 10 tap points are found in the j-th grid of image I_k , its corresponding value $g^{T_j}_{I_k}$ is 10. Obvioursly, the more frequent the touch behaviors happen in one grid, the more attentions are given to that grid by users.

Since eye fixation maps acquired from the eye-tracking device reflect real visual attention information, they are used as the ground truth for our learning algorithm. The eye fixation map is a grayscale image and each pixels value ranges from 0 to 255. The higher the value is, the more salient that pixel is. Each eye fixation map is also divided into a by b



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 1: Saliency Maps. From left to right: a) original image, b) NUSEF eye fixation map, c) our touch saliency map (gird size: image_width x image_height), d) Center saliency map, e) Itti saliency map, f) Signature saliency map, g) AIM saliency map, h) GBVS saliency map.

grids. The target real visual attention value of the *j*-th grid in image I_k: $t^j_{I_k}$ is approximated as the average of all the pixel values in the *j*-th grid. Apparently, if more pixels in one grid has high value, it indicates that grid attarcts a lot of attention.

Since different touch behaviors may contribute differently to the touch saliency value of each grid, we propose to use linear regression model to generate the touch saliency value for the *j*-th grid in image I_k in a linear function:

$$h(g^j_{I_k}) = w_0 + w_T g^{T_j}_{I_k} + w_P g^{P_j}_{I_k} + w_S g^{S_j}_{I_k} + w_D g^{D_j}_{I_k} + w_C g^{C_j}_{I_k} \quad (1)$$

w_T, w_P, w_S, w_D and w_C are the corresponding weights of the five features, which implicitly indicate correlation between each behavior and touch saliency value $h(g^j_{I_k})$.

The touch saliency learning problem is formulated as a linear regression algorithm, which learns the weight of each behavior by solving the following minimization function:

$$\min \sum_{k=1}^{m} \sum_{j=1}^{ab} \left(h(g^j_{I_k}) - t^j_{I_k} \right) \quad (2)$$

The learning framework is shown in Fig.2, it contains two stages: training and testing. During the training stage, the weight of each behavior can be learned by solving function (2), and indicates how many contributions each touch behavior makes to the touch saliency value. In the testing stage, given collected touch behavior data of a new image, its touch saliency map can be predicted with the learned weights based on formula (1). Above all, the proposed learning based approach can successfully explore the correlation between each touch behavior feature and human attention. This thus leads to a good saliency map from those touch behaviors.

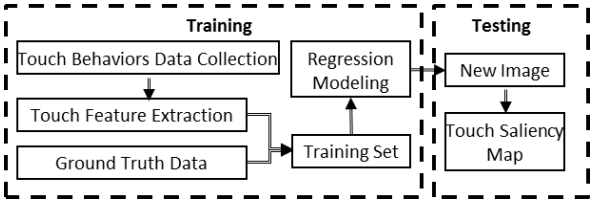


Fig. 2: Touch Saliency Learning Framework.

4. EXPERIMENTS

In our experiments, the NUSEF dataset is divided into a training set (396 images) and a testing set (50 images). After traning the model on the collected training data, the weights of features w_T, w_P, w_S, w_D and w_C are learned, whose approximate average values are 28%, 14% , 20%, 10%, and 28% respectively. These learned weights show that all the features contribute to the touch saliency, but in the different degree. Center point of screen and the tap behavior are the most important ones. Scrolling is the third important touch behavior. Pinch-zoom and Double-tap-zoom make less contribution to the visual attention information. The weights of Pinch-zoom and Double-tap-zoom are similar. This makes sense as both behaviors are used to zoom in images.

In order to evaluate the performance of our touch saliency learning from multiple touch behaviors algorithm (TSMB), we utilize two popular saliency performance evaluation metrics: AUC (Area under Curve) score and CC (Correlation Coefficients). A good saliency map should have both high AUC score (maximum value is 1) and CC score (maximum value is

Table 1: AUC and CC comparison results.

Method	Itti	GBVS	AIM	Sign.	Center	TSMB									
						10x10	14x14	18x18	22x22	26x26	30x30	40x40	50x50	60x60	WxH
AUC	0.67	0.85	0.69	0.67	0.73	0.75	0.75	0.75	0.75	0.77	0.78	0.78	0.77	0.74	0.80
CC	0.34	0.49	0.27	0.37	0.44	0.44	0.42	0.41	0.41	0.43	0.45	0.44	0.44	0.40	0.46

1). In addition, we compare the performance of our approach with other five state-of-the-arts methods on the NUSEF data set. These five state-of-the-arts include four visual saliency map generation methods, which derive saliency maps based on image visual content information (Itti Model (Itti) [7], Graph Based Visual Saliency (GBVS) [8], Attention via Information Maximization (AIM) [9], Image Signature model (Sign.) [10]), and one touch saliency generation approach (center-based touch saliency map (Center) [5]). Fig.1 shows the generated saliency maps of these methods.

In our approach, different numbers of grids are tested, which range from 10x10, 14x14, 18x18, 22x22, 26x26, 30x30, 40x40, 50x50, 60x60 to image_width x image_height. In the case of image_width x image_height(WxH), every recorded pixel in the image is chosen as one grid, the mild outliers are removed using the quartile method (lower quartile = 0*th* percentile, higher quartile = 75*th* percentile) for scroll and tap features, since most users tend to continiously scroll and accidentally tap the image on the screen.

The AUC and CC comparison result is listed in Table 1. From the results, it can be observed that: 1) our touch saliency learning algorithm TSMB outperforms the state-of-the-art touch saliency learning method (Center). The AUC value has been improved from 0.73 to 0.80 and CC value is also improved from 0.44 to 0.46. The major reason is that the center-based method only considers zoom behavior. Actually, it is found out in our study that all the touch behaviors contribute to the human attention. Tap and scroll behaviors even make more contributions than zoom does; 2) The touch saliency map generated by our algorithm has better accuracy than the saliency map derived by many complex and expensive visual-based approaches. Although multiple touch behaviors may involve noise, the generated touch saliency map still has high quality and the touch saliency learning approach is much cheaper, faster, and more efficient than visual-based approaches; 3) As the number of grids increases (the grid size decreases), the accuracy of the learned saliency map also increases. Even if the image is roughly divided into 10x10 grids, the performance is still very good. Therefore, users can freely choose the best number of grids based on their application needs. If the application has high requirement on the execution time, 10x10 is a good choice. If the accuracy is the first priority of the application, WxH should be chosen.

5. CONCLUSIONS

In this work, we conduct a quantitative and qualitative study of touch saliency learning from a variety of human touch behaviors. It is learned that different touch behaviors make different contributions to human attentions and considering more touch behaviors usually leads to a better touch saliency map. The experimental results demonstrate the proposed touch saliency learning approach can automatically generate a good saliency map from multiple human touch behaviors. Therefore, our approach will have wide application potentials where eye tracking is utilized. In the future, we will further improve the touch saliency performance by applying different learning algorithms such as classification algorithms.

6. ACKNOWLEDGEMENT

This work is in part supported by the Texas State University Research Enhancement Program (REP), Army Research Office grant W911NF-12-1-0057, and NSF CRI 1058724 to Dr. Yijuan Lu and in part supported by the NSFC 61201413, SRFDP2100060003, the Fundamental Research Funds for the Central Universities No. WK2100060007 and No. WK2100100021 to Dr. Xinmei Tian.

7. REFERENCES

[1] O. Marques, L. Mayron, G. Borba, and H. Gamba, "Using visual attention to extract regions of interest in the context of image retrieval," 2006, ACM-SE 44, pp. 638–643.

[2] S. X. Yu and D. A. Lisin, "Image compression based on visual saliency at individual scales," ISVC '09, pp. 157–166.

[3] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," ECCV'10, pp. 30–43.

[4] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," MUM '05, pp. 59–68.

[5] M. Xu, B. Ni, J. Dong, Z. Huang, M. Wang, and S. Yan, "Touch saliency," ACM MM'12, pp. 1041–1044.

[6] X. Xie, H. Liu, S. Goumaz, and W.-Y. Ma, "Learning user interest for image browsing on small-form-factor devices," CHI '05, pp. 671–680.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI'98*, vol. 20, no. 11, pp. 1254–1259.

[8] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," NIPS '06, pp. 545–552.

[9] N. Bruce and J. Tsotsos, "Saliency based on information maximization," NIPS'06, pp. 155–162.

[10] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *TPAMI'12*, vol. 34, no. 1, pp. 194–201.