HUMAN-CENTERED PICTURE SLIDESHOW PERSONALIZATION FOR MOBILE DEVICES

Cunxun Zang[1], Yu Fu[1], Jian Cheng[1], Hanqing Lu[1], Jian Ma[2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing, China
{cxzang, yfu, jcheng, luhq}@nlpr.ia.ac.cn}
[2]Nokia Research Center, Beijing,China
jian.j.ma@nokia.com

ABSTRACT

This paper presents a human-centered picture slideshow system for mobile users. In contrast to conventional ROIs (region-of-interest) detection based systems, we provide mobile users the freedom of personalizing ROIs in a convenient and effective way. Here, we import a simple human interaction, i.e., only a single click, to give a hint for users' ROIs. First, local saliency map (LSM) is generated, which considers not only multi-scale contrast, but also the self-correlation measure and central effect. Then a local fuzzy growing method is adopted to extract ROIs automatically based on LSM. Extensive experiments and user studies show the encouraging performance of the proposed system.

Index Terms— local saliency map, slideshow personalization, ROI detection, mobile devices

1. INTRODUCTION

Explosive multimedia data make our lives colorful and we can access and obtain much multimedia information anywhere and anytime with the help of mobile devices and wireless network. Meanwhile, with mobile users' accessing to more complex and larger pictures, they have higher requirement for personalization. However, due to the limited display's size of mobile devices and their small capacity of power and computation, pictures presentation on mobile devices is still an open problem. Given that picture slideshow is one of the most popular ways to present pictures [9], we focus on picture slideshow personalization for mobile devices in this paper.

Recent years, the technology of multimedia authoring achieved a great success, especially of picture slideshow. Hua et al. [7] designed an automatic picture slideshow system based on the content of the pictures and music. Later, they combined viewer's visual attention variation and integrated camera motion patterns in [8]. Zang et al. [11] proposed a multimedia messages customizing framework for mobile devices, which focused on ROI-based transition effect. Chen et al. [9] presented a tiling slideshow, in which multiple pictures sharing similar characteristics are well arranged and displayed at the same layout.

As mentioned above, most of the previous works adopted the technology of ROIs detection, which is reasonable for automatic picture slideshow generation. However, considering from a human-centered aspect, ROIs detection without human's interaction gives users very little freedom to specify their preferences, which also means that users have no capability of personalizing ROIs. For example, in the task that picture slideshow with one certain theme (or concept), perhaps only part of ROIs represent users' interest, especially for large and complex pictures. In such task, it is difficult for traditional automatic methods to detect ROIs without the help of human interaction.

Moreover, considering the inherent difficulty of understanding human interest, especially the widely existed individual differences of human interest, most of the current ROI detectors [2, 3, 4, 5, 6] may fail to provide the most attention regions of different users and two exceptions, mentioned in [12], often occur in results of ROIs detection: 1) The regions of ROI almost cover the whole image. 2) The regions that users want to focus are lost from the results. Therefore, it is not encouraged to suggest users choose ROI from the results of automatic ROIs detectors.

In this paper, we present a convenient and effective way to personalize picture slideshow for mobile users. Here, a simple human interaction, i.e., only one single click, named point-of-interest, is provided for users to specify their ROIs. To obtain user's ROIs, local saliency map is firstly generated, where multi-scale color contrast, self-correlation measure and central effect are considered. Then a local fuzzy growing method [3] is adopted to extract ROIs, which is finally specified by a rectangle. Compared with existing automatic systems, the proposed system has the following advantages. 1) A simple human interaction is imported to indicate users' preferences. 2) A Self-correlation measure is proposed to provide the appearance guide for ROIs detection. 3) Personalized and effective ROIs results are obtained whose arbitrariness and time cost is significantly reduced.

2. SYSTEM DESIGN

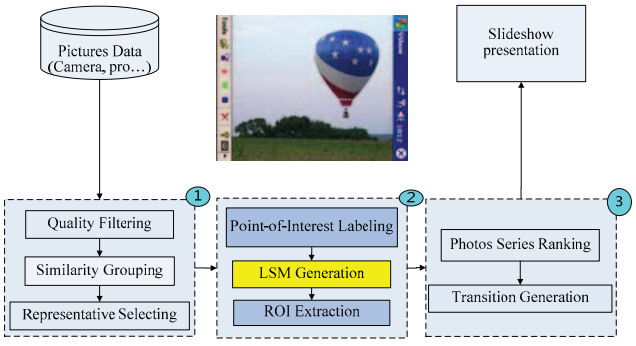


Figure 1. System architecture of our approch.

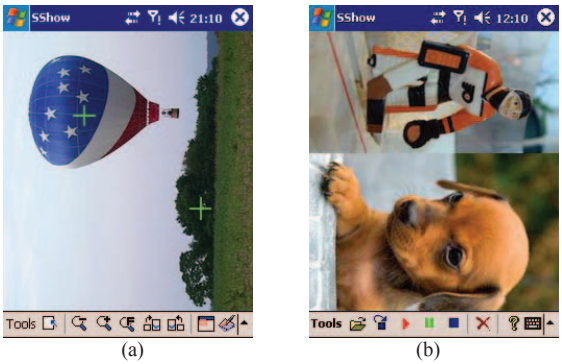
The proposed system architecture is presented in Fig. 1, which comprises three components, i.e., content-based visual clustering, ROIs personalization and picture series generation. The workflow of our system is introduced as follows:

1. *Content-based visual clustering:*
   1) A quality filter [7] is firstly used to remove poor visual pictures from the input picture data.
   2) Then pictures' grouping and selecting are processed to remove pictures which are too similar. Here, the best-first merging method [10] is employed.
2. *ROIs personalization:*
   1) To specify users' personalized ROIs, the positions corresponding to users' click, named points-of-interest, are firstly labeled, whose process is illuminated in Section 2.1.
   2) Then local saliency map (LSM) is automatically calculated, whose methods is described in Section 2.2.
   3) Based on LSM, ROIs are extracted by a local fuzzy growing method [3]. More details are in Section 2.3.
3. *Picture series generation:*
   1) Based on ROIs, pictures are ranked in a reasonable and pleasing order using the optimization method of [12].
   2) After that, the method of ROI based transition effect in [12] is employed to help mobile users more effectively keep attention regions in sight.

Next, we will describe the 2nd component of the proposed system step by step, which is also the main novelty of our work.

2.1. Point-of-Interest Labeling

In our system, a convenient interaction way is provided for mobile users to specify their ROIs of image content for certain purposes (theme or concept), i.e., only a single click on the display screen. Here, the position of click is named point-of-interest, which is also referred in [13] and usually lies nearby the center of the attention region. In our system, multiple point-of-interests are allowed in one picture, and each of them marks one attention region. As shown in Fig. 2(a), one green cross denotes one point-of-interest. To avoid invalid operations, too closer points are merged into one point, whose position is calculated in average.

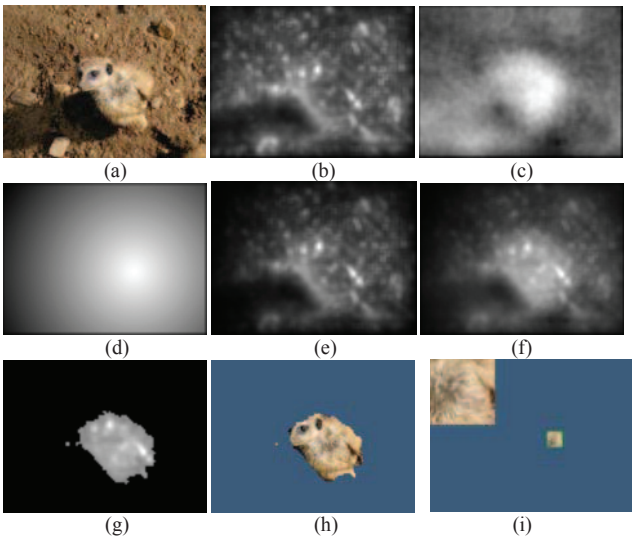


(a) (b)

Figure 2. (a) UI for points-of-interest labeling. (b) UI for slideshow player.

2.2. Local Saliency Map Generation

Different from traditional saliency map, the concept of local saliency map [12] can be considered as a dynamic saliency map with user's guidance. While only image content is known for traditional saliency map [2]; the additional information, i.e., user's guidance, as well as the corresponding appearance prior, is provided for local saliency map. In this paper, we propose a more effective method for local saliency map generation, where the self-correlation measure is integrated effectively, whose efficiency is verified by experiments in Section 3.1.



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

Figure 3. (a) Original image. (b) color contrast map. (c) self-correlation map. (d)central effect map. (e) LSM by [12]. (f) LSM by the proposed method. (g) the result of local fuzzy growing. (h) pixels corresponding with (g). (i) image patch used for self-correlation map.

To generate local saliency map, here, three factors are considered from different aspects:

1) *Contrast measure*. Results from vision science, such as [1] suggest that saliency can be measured by low-level feature contrast. This serves as the theoretical basis of all the existing methods [2, 3, 4, 5, 6]. Multi-scale color contrast measure is employed, which has been proved to be effective in practice [5].

2) *Central effect*. The position factor is useful for local saliency map and it is reasonable that pixels with smaller distance from the point-of-interest have higher possibility to belong to attention region. In other words, the attention of users decreases with the distance between current position and the point-of-interest increases gradually. As the shape prior of attention region is unknown, we use isotropic center effect map in our work.

3) *Self-correlation measure*. In this paper, besides of the first two factors above, self-correlation information in color and texture is also considered, which is an important factor widely existed in its universal existence in nature objects. In fact, self-correlation has been employed in image segmentation [13, 14], but to our best knowledge, rarely no one has used it in saliency map. As a pixel's self similarities descriptor, Self-correlation measure can provide an appearance prior for saliency map and obtain a more semantic and specific region extract by merging the first two factors An example of self-correlation map is shown in Fig. 3(c). Note that, we mainly focused on to provide one of the most salient region near the attention point, and the case that many segments with two complex textures and colors in a large scope can be handled by multiple user interactions.

In the following, we outline the proposed method for local saliency map generation in details:

Step 1: Image resizing. The image is firstly resized to a uniform size with its aspect ratio unchanged [3], which has two advantages.

1) All images are considered in the same scale. 2) The computational complexity is effectively reduced.

Step 2: Color space transformation. As CIELAB space is consistent well with human color perception system, the resized image is transformed from RGB space to CIELAB space.

Step 3: Multi-scale contrast map generation. Considering that the size of salient object is unknown, color contrast is computed at multiple scales [5]. Here, a simple multi-scale contrast measure is defined as a linear combination of contrasts in the Gaussian image pyramid:

$$f(x,M) = \sum_{l=1}^{L} \sum_{x' \in S} \| I^l(x) - I^l(x') \|^2 \quad (1)$$

where I^l is the l level image M in the pyramid and the number of pyramid levels L is 3. S is a 5×5 window. x is the position of (i, j). The multi-scale contrast map $f(x,M)$ is normalized to a fixed range [0, 1]. An example is shown in Fig. 3(b).

Step 4: Self-correlation map generation. Here, an assumption in [14] is also made that the three color channels in CIELAB space are statistically independently. For each pixel in image, a 21x21 image patch, shown in Fig. 3(i), is used to describe color distribution in three channels of Lab respectively. To measure the correlation between two pixels, normalized symmetric mutual information [14] is adopted. For each channel, 128-bins histogram is used to build color probability. The self-correlation map is created as below:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (2)$$

$$NI(X,Y) \approx \frac{2 \cdot I(X,Y)}{H(X) + H(Y)} \quad (3)$$

$$sc(x,M) = NI(X_L, C) \cdot NI(X_a, C) \cdot NI(X_b, C) \quad (4)$$

where X, Y, C denote color distribution at the position of x, y and the point-of-interest, respectively. $H(X)$ is the entropy of X. X_L, X_a, X_b denote three channels of X, respectively. The self-correlation map is normalized to a fixed range [0, 1]. An example is shown in Fig. 3(c).

Step 5: Central effect map generation. Different from [12], central effect is defined as a simple linear function.

$$ce(x,M) = 1 - \frac{D_x}{D_{\max}} \quad (5)$$

where D_x is the distance from x to the point-of-interest and $D_{\max}$ is the max distance to the point-of-interest. An example is shown in Fig. 3(d).

Step 6: Local saliency map generation. Here, local saliency map is considered as a simple linear combination of contrast and self-correlation maps under central effect.

$$lsm(x,M) = N\left(\alpha \cdot f(x,M) + \beta \cdot sc(x,M)\right) \cdot ce(x,M) \quad (6)$$

where $N(\bullet)$ denotes the normalization operation. In our system, $\alpha = 0.6, \beta = 0.4$. The local saliency map is normalized to a fixed range [0, 1]. An example is shown in Fig. 3(f).

2.3. ROI Extraction

After LSM generation, ROI is extracted based on the method of fuzzy growing [3]. Different from [3], only one ROI is extracted from one local saliency map. Hence, we regard our ROI extractor as a local fuzzy growing method. Some results of our extractor are shown in Fig. 4, where the first column are results by different detectors, the second are local saliency map results and the third are local fuzzy growing results.

As shown in Fig. 4(d), the green cross is the point-of-interest marked by user's single click. The red and blue regions are results of the attention view and attention areas extractors in [3], respectively. The ultramarine and yellow regions are results of [12] and our proposed method, respectively. Note that the blue region is not always marked, that is because the attention areas detector in [3] doesn't always output reasonable results. No blue region is marked in case of results of the attention areas detector [3] cover almost the full picture. Notice that in Fig. 4(j, m), different points-of-interest are marked and our method works well for such personalized ROIs.

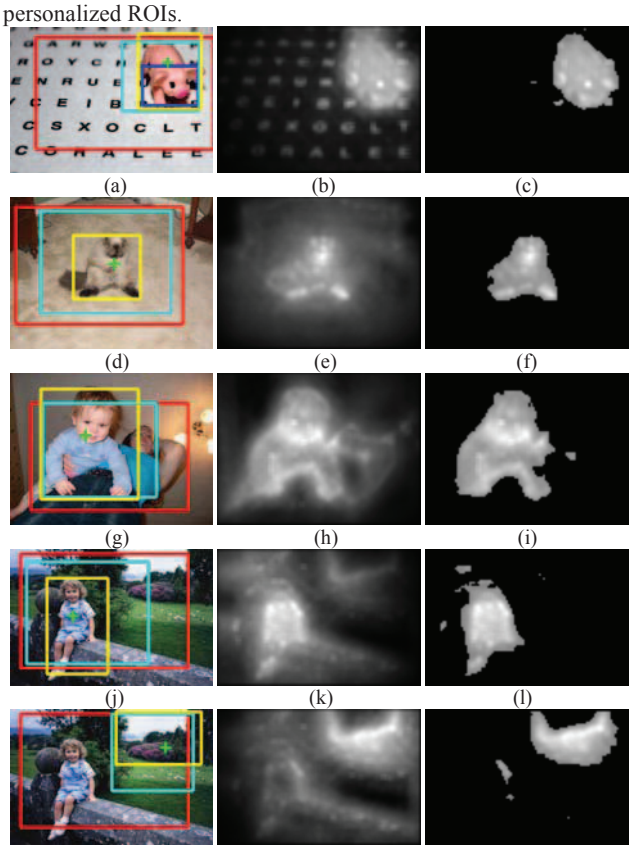


Figure 4. (a, d, g, j, m) original pictures. The green cross is the point-of-interest marked by user's single click. The red, blue, ultramarine and yellow regions are results of the attention view and attention areas extractors in [3], attention view detector in [12] and our proposed method, respectively. (b, e, h, k, n) local saliency map. (c, f, i, l, o) local fuzzy growing.

3. EVALUATION

Our prototype system is developed on the hardware of Dopod 696 (Pocket-PC), and running on the operation system of Microsoft Windows CE.Net. UI for points-of-interest labeling and slideshow player are shown in Fig.2 (a, b) respectively.

3.1. LSM-based ROI Detection

We tested the performance of our proposed ROIs detector. The experiment data consists of 500 selected images from MSRA Salient Object Database [15]. 50 pictures are selected from each of

10 file folders [14] and pictures with multiple ROIs have priority of selection. 16 computer-science students are involved help finishing the experiment, who are familiar with the operations on mobile devices. They marked all the ground truth of attention regions. For one picture, one or multiple ROIs are labeled in our experiment. For MSRA Salient Object Database, the original ground truth (only one ROI for one picture) is reserved. Considering that traditional attention region detectors are calculated without user's guidance, for the purpose of fair comparison, we compare our proposed detector with that of [12]. But we also record the results of an effective traditional attention region detector [3] (part of results are shown in Fig. 4, marked by red and blue regions) to provide a valid reference.

To evaluate the precision of attention region detection, an area-based MI (mutuality information) is defined as

$$MI_{area}(i,j) = \frac{Area_{av}i \cap Area_{av}j}{Area_{av}i \cup Area_{av}j} \quad (7)$$

The average MI values of two detectors are shown in Fig. 5. We find that the performance of our proposed detector is significantly higher than that of [12], which verifies the efficiency of the self-correlation measure integrated in our method.

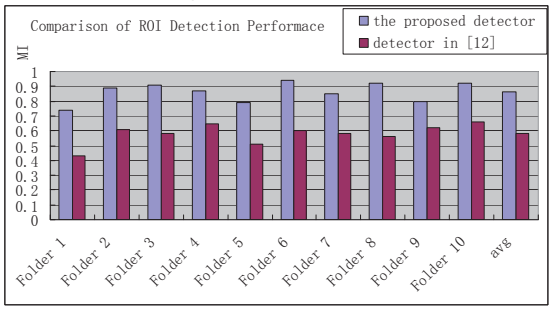


Figure 5. Performance Comparison of ROIs Detection.

3.2. User Study

A user study is conducted to evaluate the satisfaction of our system. Here, we focus on examining the validity of the personalized ROIs detection component described in Section 2. Hence the task is designed to specify ROIs with the certain theme: animal. Two methods are compared in the user study, the first is our proposed method and the latter is manual method. In our method, users only need to mark points-of-interest and then LSM generation and ROIs extraction are calculated automatically. In manual method, users need draw rectangles in pictures to specify ROIs directly, which are also regarded as the ground truth. In both methods, the basic functions of the traditional image browser (e.g., zoom in, zoom out, pan the whole image, zoom in one selected region, etc.) are provided. The same pictures in Section 3.1 are used and the same 16 students are involved.

Table 1. Average number of actions per picture

Action	*Scroll*	*Zoom in*	*Zoom out*	*Click*	*Drag-drop*
Manual	*0.31*	*0.44*	*0.30*	*0.37*	*1.98*
Ours	*0.23*	*0.52*	*0.27*	*1.62*	*0.25*

For each user, we record their browse log. The result of average number of actions per image is shown in Table 1, where the action of Drag-drop is defined as a series of successive actions: click, move and up from the screen. As shown in Table 1, both methods require the similar number of actions. However, since the number of the drag-drop action is reduced obviously, our method can reduce the complexity of actions. Notice that there are less zoom in, zoom out and scroll actions in our method. The performance of our ROIs detector is required to be improved in the further, which is also our future work.

4. CONCLUSIONS

This paper presents a human-centered picture slideshow system for mobile users. Compared with traditional ROIs (region-of-interest) detection based systems, we provide mobile users the freedom of personalizing ROIs in a convenient and effective way. Here, we import a simple human interaction, i.e., only a single click, to give a hint for users' ROIs. First, local saliency map is generated, which considers not only multi-scale contrast, but also the self-correlation measure and central effect. Then a local fuzzy growing method is adopted to extract ROIs automatically based on local saliency map. Extensive experiments and user studies show the encouraging performance of the proposed system.

5. CACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 60833006), the Natural Science Foundation of Beijing (Grant No. 4072025) and the National High Technology Research and Development Program (863) (Grant No. 2006AA01Z117).

6. REFERENCES

[1] H.C. Nothdurft, "Salience from feature contrast: additivity across dimensions," Vision Research, vol. 40, no. 11-12, pp. 1183–1201, 2000.

[2] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," PAMI, 1998.

[3] Y.F. Ma, and H.J. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing," ACM Multimedia 2003.

[4] L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, "A visual attention model for adapting images on small displays," ACM Multimedia Systems Journal, 2003.

[5] T. Liu, J. Sun, N.N. Zheng, X.O. Tang and H.Y. Shum. "Learning to Detect A Salient Object," CVPR 2007.

[6] X.D. Hou, and L.Q. Zhang, "Saliency Detection: A Spectral Residual Approach," CVPR 2007.

[7] X.S. Hua, L. Lu, and H.J. Zhang, "Content-Based Photograph Slide Show with Incidental Music," Proc. of ISCAS 2003. Vol. II, pp. 648 -651, 2003.

[8] X.S. Hua, L. Lu, and H.J. Zhang, "Automatically Converting Photographic Series into Video," ACM Multimedia 2003.

[9] J. Chen, W. Chu, J. Kuo, C. Weng, and J. Wu, "Tiling Slideshow," ACM Multimedia, 2006.

[10] J. Platt, "Auto Album: Clustering Digital Photographs using Probabilistic Model Merging," IEEE Workshop on Content-Based Access to Image and Video Libraries 2000.

[11] C.X. Zang, Q.S. Liu, H.Q. Lu, K.Q. Wang, "A New Multimedia Message Customizing Framework for Mobile Devices," ICME 2007.

[12] C.X. Zang, Q.S. Liu, J. Cheng, and H.Q. Lu, "Human-centered image navigation on mobile devices," ICME 2008.

[13] S. Bagon, O. Boiman, and M. Irani, "What is a Good Image Segment? A Unified Approach to Segment Extraction," ECCV 2008.

[14] http://www.lans.ece.utexas.edu/~strehl/diss/node107.html

[15] http://research.microsoft.com/~jiansun/SalientObject/salient_object.html