A NEW MULTIMEDIA MESSAGE CUSTOMIZING FRAMEWORK FOR MOBILE DEVICES

Cunxun Zang[1], Qingshan Liu[1], Hanqing Lu[1], Kongqiao Wang[2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing, China
{cxzang, qsliu, luhq}@nlpr.ia.ac.cn}
[2]Nokia Research Center, No.11 He Ping Li Dong Jie, Nokia House 1,Beijing 100013,China
kongqiao.wang@nokia.com

ABSTRACT

In this paper, we present a novel framework to customize multimedia messages for mobile users. The goal is to generate a video message from a series of pictures. The framework includes visual attention view detection, image grouping, image ranking, and slideshow generation. Considering the limitation of mobile device, we use a simple color feature based attention model to detect interesting regions of the images. We group the images, and rank them based on the attention view similarities. Finally a human perception based slideshow is designed to keep the mobile users' eye on attention regions efficiently. In addition, a short music is selected to match the video message. Extensive experiments and user studies show the promising performance of the proposed system.

1. INTRODUCTION

It is known that explosive multimedia data make our lives colorful. Especially with the help of smart mobile devices and wireless network, we can connect and obtain all kinds of multimedia information anywhere and anytime. Multimedia message service (MMS) is a revolutionary successor to short messaging service (SMS), and becoming a new and important standard for messaging services on mobile phones [2]. In this paper, we will focus on how to customize personal MMS for mobile users.

In previous work, there are some studies on MMS analysis. Ling et al [1] analyzed grounded genres in multimedia messaging. Majid et al. [2] proposed an algorithm to find the optimal temporary storage size for MMS system. Vatsa et al. [3] focused on the importance and beneficial role of media transformation. Zhong et al. [4] considered energy-efficient design and construct a hierarchical system for users to access multimedia content. However, few studies focus on customizing multimedia messages on mobile devices. Customizing multimedia message on mobile devices is a challenge work, because mobile device have several limitations, such as, computation capacity, screen display size, network bandwidth, battery lifetime, and user interface.

The technology of video authoring achieved a great success in recent years. Hua et al. [7] designed a 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 for picture slideshow in [8]. Chen et al. [9] presented a tiling slideshow, in which multiple pictures sharing similar characteristics are well arranged and displayed at the same layout. But these techniques [7,8,9] cannot be directly transferred to mobile devices due to high computation cost. In addition, the layout style of tiling slideshow is not suitable for mobile devices with small screen display. In [10], a visual attention model was developed for adapting images on small displays, and a similar scheme was presented in [11] for browsing large pictures automatically on mobile devices. However, both of them only discussed to browse images on small displays.

In this paper, we present a novel and practical framework to customize multimedia video messages for mobile users. The framework includes visual attention view detection, image grouping, image ranking, and slideshow generation. After quality filtering, we group pictures in limited-number clusters, because it is forbidden by a MMS to contain too many pictures [2]. A simple color feature based attention model is designed to detect interesting regions of the images similar to [6]. The images are ranked with the attention view similarities. Finally a human perception based slideshow is designed to keep the mobile users' eye on attention regions efficiently. In addition, the video message is matching with a short music. Extensive experiments and user studies show the promising performance of the proposed system.

2. SYSTEM DESIGN

The proposed system architecture is presented in Figure 1, which comprises four core components, i.e., content-based visual clustering, attention view detection, picture series ranking, and transition-effect generation.

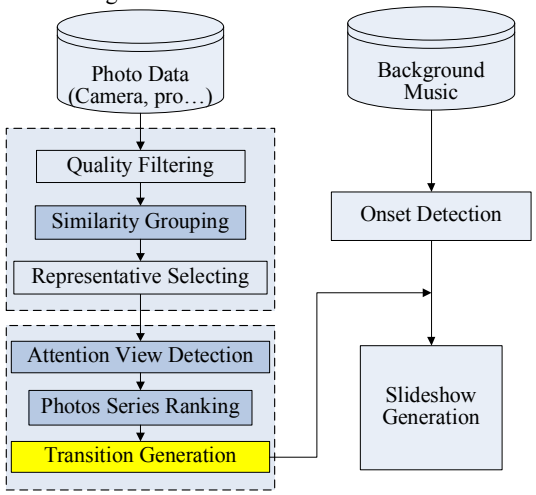

Figure 1. System architecture of our approch.

The system workflow can be formulated as:

1) A quality filter is used to remove poor visual pictures from the input picture data, which contain pictures captured by mobile cameras or ones from PC or internet.

2) Pictures' Grouping and selecting are processed to remove too similar pictures.
3) A simplified attention view detector is employed to search attention view center and region.
4) Pictures are ranked based on the information detected in (3).
5) Attention-based transition-effects are generated, whose details are described in section 2.5.
6) The onsets information of background music is detected on PC by the method in [7] and stored into xml files.
7) A slideshow is generated based on information above, where pictures are transited in form of our proposed attention-based transition effects and accompanied with the pace of background music.

2.1. Quality Filtering

In our system, picture data have two sources: ones captured by cameras on mobile devices and ones downloaded from PC or internet. Since most of captured ones by cameras are raw, a quality filter is obviously necessary. Considering the limitation of computing capability, the quality filter in our system is similar to [7], and only considers two cases, i.e., homogenous and under or over exposed pictures.

2.2. Pictures Grouping and Selecting

To group pictures, the best-first merging method in [12] is employed in our system. HSV color histogram is adopted to calculate the distances between two pictures. The grouping algorithm is described briefly as follows:

1) Input picture sequence:
$GroupList = \{Group_i = ph_i, 1 < i < M\}$,
$GroupNum = M$. (M is the number of input pictures.)
2) If $\forall L_i >= L_{\min}, 1<i<GroupNum$ and $GroupNum \le \min(M, N_{\max})$, then exit, where $N_{\max}$ is the whole number of pictures; otherwise compute the distances between every pair groups in $GroupList$ and find out the best-merged (closed) two groups: $Group_i$ and $Group_j$, merge $Group_i$ and $Group_j$ into $Group_{\min(i,j)}$

2.3. Attention View Detection

The attention view detection is helpful to understand semantics of the pictures in a sense. However, as mentioned in introduction, it is impractical to detect attention view on mobile devices with a complex attention model due to its high computing complexity. To solve this problem, we propose a simplified scheme based on that in [6] to realize attention view detection on mobile devices, which can keep an appropriate performance.

The process of attention view detection is summarized as follows:

1) The original picture is resized to 320x240 pixels.
2) Convert the resized picture from RGB space to HSV space.
3) Color quantization. The peer group filtering (PGF) method [13] adopted in [6] can remove impulse noise effectively and smooth color images without blurring edges and details, but it has a relative high computational complexity for mobile devices. To archive a reasonable trade off between the performance and computational complexity, a uniform quantization method is employed to replace PGF, in which every component of HSV space is divided into 16 bins in average.
4) The resized picture is down-sampled into 40x30 pixels.
5) Saliency map is constructed as in [6].
6) Based on the saliency map, an attention view is defined as a region with $viewcenter = (x_c, y_c)$ and $viewsize = (W, H)$, which can be calculated as follows:

$$x_c = \sum_{i=0}^{M-1} \left(\frac{1}{CM} \sum_{j=0}^{N-1} C_{i,j} \cdot i\right) \quad (2)$$

$$y_c = \sum_{j=0}^{N-1} \left(\frac{1}{CM} \sum_{i=0}^{M-1} C_{i,j} \cdot j\right) \quad (3)$$

$$W = 2\alpha \cdot \sum_{i=0}^{M-1} \left(\frac{1}{CM} \sum_{j=0}^{N-1} C_{i,j} \cdot |i - x_c|\right) \quad (4)$$

$$H = 2\beta \cdot \sum_{j=0}^{N-1} \left(\frac{1}{CM} \sum_{i=0}^{M-1} C_{i,j} \cdot |j - y_c|\right) \quad (5)$$

$$CM = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} C_{i,j} \quad (6)$$

In our system, $\alpha = 1.8, \beta = 1.8, M = 320, N = 240$.

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

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

As shown in Figure 2(b), the blue, yellow and red regions are the attention views detected by [6], detected by our method, and marked manually for ground truth respectively. The MI values of [6] and our method are 0.622 and 0.619 respectively. We can see that the performance of the proposed simplified attention model is acceptable.

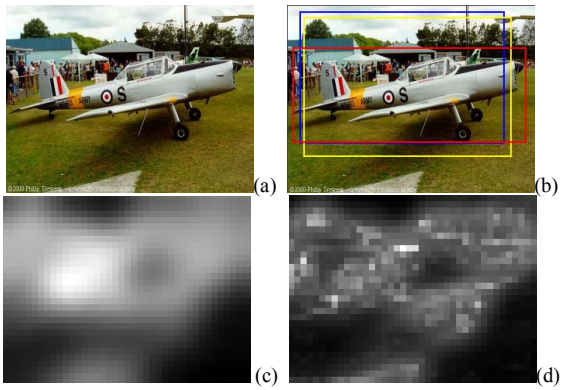

Figure 2. Contrast-base attention view analysis. (a) original picture (b) attention results. (c,d) saliency map by [6] and our simplied method.

2.4. Picture series ranking

To propose a more reasonable and pleasing slideshow, we consider the following principles that are suggested by several users for photographs transformations:

a) Attention view of new pictures should be shown firstly and keep in insight as long as possible.
b) Attention view of adjacent pictures should be scattered, i.e., the distance of attention view center between

adjacent pictures should be as large as possible, which seems more aesthetic.

Therefore, we simply rank the picture sequence based on the attention view distance:

$$D_{av}(i,j) = \alpha \cdot D_{center}(i,j) + \beta \cdot MI(i,j) \quad (8)$$

$$D_{center}(i,j) = \| av_{center}i - av_{center}j \| \quad (9)$$

In our system, α=0.9,β=0.1 .

Since the maximum number of photographs in our system is limited to 6, the full search method is used to find the ranking sequence:

$$\Theta^{*} = \arg\max\left(\sum_{i \in \Theta} D_{av}(i,i+1) \right) \quad (10)$$

2.5. Transition Effects Generation

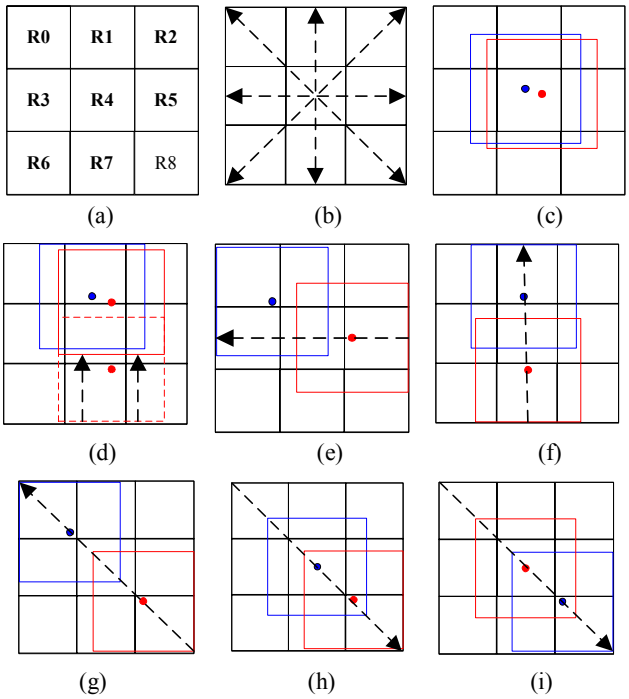

Figure 3. Cases of adjacent pictures' attentive views and forwrad directions for slide and push transition effects.

Because the display screen is small for mobile devices, it cannot list several images in one slide. In the proposed system, we just show one picture in one slide, and we focus on designing the transition effects between two slides, which is different from most previous slideshow schemes [7,8,9]. We consider the attention views of the adjacent pictures and generate transitions to fulfill the principles described in section 2.4 for maximum possibility.

First, a picture is divided into 9 regions (from R_0 to R_8) in average shown in Figure 3 (a). Three types of common transition effects, i.e., slide, push, and fade effects are selected for candidates. For both the slide and push transition effects, 8 forward directions are defined in Figure 3(b). In Figure 3 from (c) to (i), the blue region with a blue center point denotes the attention view (AV) of the prior picture (Pic_i) and the red one is marked as that of the next one (Pic_{i+1}) respectively, whose region centers of AV are denoted as Rc_i and Rc_{i+1}. To select a more reasonable transition effect, different cases of Rc_i and Rc_{i+1} are analyzed as follows:

1. Rc_i and Rc_{i+1} in the same region.
   - In the Center Region R_4: The fade effect is chosen simply.
   - In the Center Region R_n ($n \neq 4$): The push effect is selected, whose push direction is from the opposite direction of R_n to that of R_n. As shown in Figure 3 (d), Pic_{i+1} is pushed from the bottom to the top gradually.
2. Rc_i and Rc_{i+1} in different regions. The moving effect of slide transition is based on the positions of Rc_i and Rc_{i+1}:
   - One is in the Center Region R_4 and the other is in one of the Corner Regions (R_0, R_2, R_6 and R_8). A tilted moving direction is selected from the Center Region to the Corner Region shown in Figure 3 (h and i).
   - Both in the opposite pare Corner Regions (R_0 and R_8 or R_2 and R_6). The direction is from Pic_{i+1} to Pic_i shown in Figure 3 (g).
   - Other cases. The direction is horizontal or vertical, which is also from Pic_{i+1} to Pic_i shown in Figure 3 (e and f).

2.6. Onset Detection

In our system, pictures transition happens at the beat positions of the background music, which is detected on PC by the method in [7] and is stored into xml files.

3. EVALUATION

We develop our system on the Pocket-PC Dopod 696, which is running on Microsoft Windows CE.Net as the operation system. As shown in Figure 4, (a) is the UI of selecting picture candidates and (b) is that for slideshow, in which the push transition effect is selected and the yellow arrow represents the motion direction.

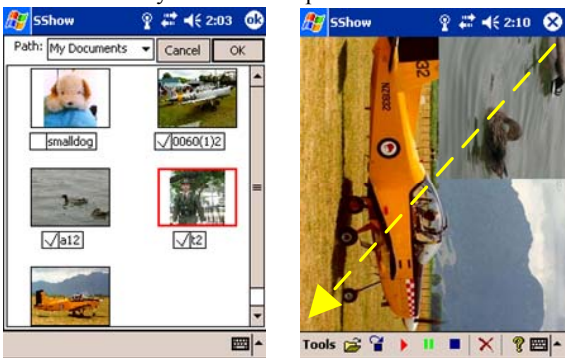

Figure 4. UI of our system on dopod 696. (a) UI of selecting picture candidats. (b) UI of attention-based slideshow.

3.1. Attention View Detection Performance

To test the performance of our simplified attention view detector, 40 photos are selected from the benchmark Caltech 101 dataset [14] and 40 ones are captured by mobile cameras. The average values of the time cost are 5.18s and 8.32s for our simplified detector and [6] respectively, because we replaced the non-uniform quantizing with a simple uniform one, down-sampled the resolution of pictures, and decreased the degree of refinement in color space. The MI values are 0.713 and 0.695 for our simplified detector and [6] respectively. It demonstrates our detector keeps an appropriate performance. Some results from Caltech 101 dataset [14] are shown in Figure 4.

3.2. User Study

A user study was conducted to evaluate the satisfaction of our system on customizing multimedia messages. We compare our system with the original video message system in Dopod 696, which simply slides pictures with normal transitions. 16 computer-science students are involved, who are familiar with the operations on mobile devices. Firstly, they were taught to edit multimedia messages by both tools. Then we collected some pictures from the benchmark Caltech 101 dataset [14] and ones captured by the camera on Dopod 696. For each user, we fielded a similar questionnaire same as [9], which concerned the five aspects of our system to get their feedbacks and should be answered on a five-scales scores from 1 to 10 (high score means better satisfaction.

1) Fun: Is it a funny tool on customizing multimedia messages?
2) Effectiveness: Is it enough effective to multimedia messages?
3) Practicality: Is it practical to operate it?
4) Atmosphere: How do you feel the audiovisual effects?
5) Atmosphere: How do you feel the photographs transition?

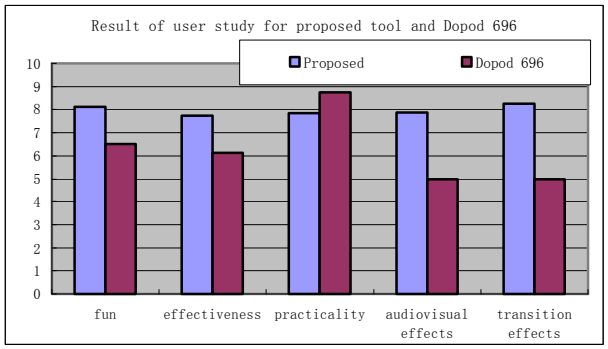

Figure 5. Result of user study for our system and Dopod 696.

The results of five aspects tests are illustrated in Figure 5. Generally, our proposed system has significantly better satisfactions than the original multimedia system in Dopod 696. Especially, the effect of our photographs transition based on the attention view has been certificated by these 16 users. Though we also observe that the practicality of our system is lower obviously than that of Dopod 696, it is still tolerable if we consider that many more computing processes have been done to analyze photographs content and rank photographs. And we will attempt to improve the practicality of our system in the future work.

4. CONCLUSIONS

This paper presents a novel framework that customizes multimedia messages for mobile devices. The proposed framework is composed of content-based clustering, attention view detection, picture series ranking and slideshow transition-effect generation. We group pictures in limited-number clusters based on visual similarity, rank clusters' representatives based on the maximum total distance of attention view, and slide them with our proposed transition-effect based on attention view to help mobile users more effectively keep attention regions in sight. Experiments are presented with promising results on validating the practicality and effectiveness of our system.

4. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 60475010 and 60121302), the Natural Science Foundation of Beijing (Grant No. 4072025) and the National High Technology Research and Development Program (863) (Grant No. 2006AA01Z315).

5. REFERENCES

[1] Ling, R. and T. Julsrud. "Grounded genres in multimedia messaging." Pp. 329 - 338 in A sense of place: The global and the local in mobile communicaiton, edited by K. Nyiri. Vienna: Passagen Verlag, 2005.

[2] Majid Ghaderi, Srinivasan Keshav, "Multimedia Messaging Service: System Description and Performance Analysis," pp. 198-205, WICON 2005.

[3] Vatsa R., Kumar V., "Role of Media Transformation in Multimedia Messaging", Pp. 258 – 262, ICPWC 2005.

[4] Zhong, L., Wei, B., and Sinclair, M. J. 2006. SMERT: energy-efficient design of a multimedia messaging system for mobile devices. In Proceedings of the 43rd Annual Conference on Design Automation, ACM Press, 2006

[5] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998.

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

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

[8] Hua, X.S., Lie LU, and Hong-Jiang ZHANG, "Automatically Converting Photographic Series into Video," ACM *Multimedia* 2003.

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

[10] 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.

[11] Liu, H., Xie X., et al, "Automatic Browsing of Large Pictures on Mobile Devices", ACM *Multimedia* 2003.

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

[13] Y. Deng, C. Kenney, *et al.*, "Peer group filtering and perceptual color image quantization," *Proc. of* IEEE International Symposium on Circuits and Systems, Vol.4, p.21-24, 1999.

[14] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

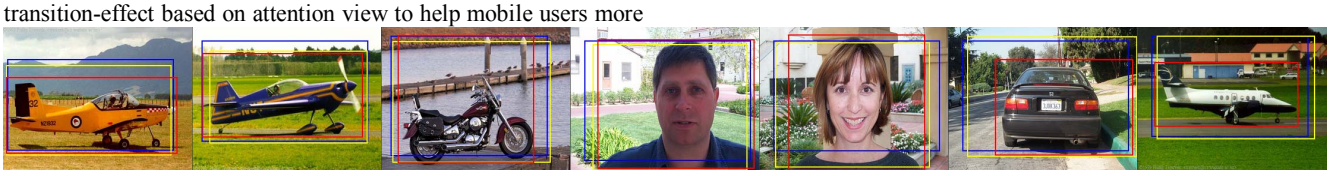

Figure 6. Some results of attention view detection in Caltech 101.