EFFICIENT SALIENCY-BASED REPURPOSING METHOD

O. Le Meur, X. Castellan

THOMSON R&D
1 Avenue Belle Fontaine
35511 Cesson-Sevigne, France

P. Le Callet, D. Barba

IRCCyN UMR 6597 CNRS
Ecole Polytechnique de l'Universite de Nantes
rue Christian Pauc, La Chantrerie, 44306 Nantes, France

ABSTRACT

Images play a very relevant role in our daily life. People now can easily shoot and share pictures thanks to the exponential growth of the portable medias, such as digital cameras, mobile phone... As the display size of those devices is relatively small, browsing large pictures remains difficult. Content repurposing is an elegant solution to deal with this problem. It consists in cropping the images in order to display only the most interesting parts of the picture. A new algorithm is proposed in this paper; the experiments described herein, leading to a qualitative and a quantitative assessment, show that the proposed solution outperforms the conventional method.

1. INTRODUCTION

Considering both the variety of display devices (PDA, mobile phones...) and the huge amount of picture/video format, a cost-effective method is primordial to efficiently manage this diversity. Content repurposing is an elegant solution to deal with this problem. Picture/video material coming from one particular network or well suited to a particular display's size are adapted to meet the constraints of another one. As manually repurposing of images is in most cases unconceivable, the usual way to display a large image on small screen is to dramatically downsample the picture. The reduced picture is commonly called thumbnail. Nevertheless, such a conventional approach often yields unworkable pictures because important objects of the scene could be unrecognizable.
In order to provide thumbnails of optimum quality for the targeted device, it is both necessary to identify important regions of interest and to compute the reduced picture centered on these parts. Recents methods rely on this rationale and use a cropping method discarding the less important parts of the image. Cropping-based methods have been proposed by a number of authors [1, 2]. In such approaches, the thumbnail is a subset of the picture, meaning that the context can be lost. A recent method [3] uses a non-linear image warping function in order to keep the background, e.g. the less important aspects of the image. Nevertheless, the deformation can be bothersome in numerous cases.
The major problem in the cropping-based method is obviously not the cropping but rather the identification of the most interesting parts of the picture. A visual attention model is a good solution to overcome this difficulty [1].
In this paper, a non-linear saliency-based thumbnail determination is proposed. The proposed method, similar in spirit to that of [1, 2], uses the results coming from either an eye tracking apparatus or a visual attention model. Section II describes the computation of both the saliency map stemming from eye tracking experiments and the predicted saliency map obtained by a computational model of the selective visual attention [4, 5]. Section III describes the automatic cropping algorithm. In section IV, a qualitative and quantitative comparison is done taking into account conventional thumbnails, thumbnails based on eye tracking experiments and thumbnails based on a predicted saliency map.

2. EXPERIMENTAL AND PREDICTED SALIENCY MAPS

2.1. Experimental saliency maps

In order to track and record real observers eye movements, experiments have been performed with a dual-Purkinje eye tracker from *Cambridge Research Corporation*. The eyetracker is mounted on a rigid EyeLock headrest that incorporates an infrared camera, an infrared mirror and two infrared illumination sources. To obtain accurate data regarding the diameter of the subjects's pupil a calibration procedure is needed. The calibration requires the subject to view a number of screen targets from a known distance. Once the calibration procedure is complete and a stimulus has been loaded, the system is able to track a subject's eye movement. The camera recorded a close-up image of the eye. Video was processed in real-time to extract the spatial location of the eye position. Both Purkinje reflections are used to calculate the location. The guaranteed sampling frequency is $50\ Hz$ and the accuracy is about 0.5 degree. Forty subjects participated in the experiments. They came from both the university of Nantes and from Thomson R&D Rennes. All had normal or corrected to normal vision. All were inexperienced observers (not expert in video processing) and naive to the experiment. Before each trial, the subject's head was positioned so that their chin rested on the

Fig. 1. (a) original picture; (b) experimental saliency maps; (c) predicted saliency maps.

chin-rest and their forehead rested against the head-strap. The height of the chin-rest and head-strap was adjusted so that the subject was comfortable and their eye level with the center of the presentation display.
Twenty five pictures of various contents have been selected. Each picture was presented to subjects in a free-viewing task. Subjects were instructed to "look around the image". For each observer, the calibration of the eye tracker was intermittently redone between two pictures as required. Experiments were conducted in normalized conditions (ITU-R BT 500-10) at viewing distance of five times the TV monitor (800) height. The free viewing condition is mandatory, in order to lessen the top-down effects. It is required that visual attention is mainly driven by the low level visual features.
From the collected data, a fixation map is computed for each observer. It encodes the saliency degree of each spatial location of the picture. This kind of map is often compared to a landscape map consisting of peaks and valleys. A peak, indicating the number of fixations, represents the observer's regions of interest. A saliency map SM^k for an observer k is given by

$$SM^k(x,y) = \sum_{j=1}^{NbData} \Delta(x - x_j, y - y_j) \qquad (1)$$

$NbData$ is the number of data collected by the eye tracker apparatus.
To determine the most visually important regions, all the fixation maps are merged yielding to an average fixation map SM:

$$SM(x,y) = \frac{1}{N} \sum_{k=1}^{N} SM^k(x,y) \qquad (2)$$

N is the number of observer. The average saliency map encodes the most attractive part of a picture when a large panel of observers is considered. Finally, the average saliency map is smoothed with a 2D Gaussian filter given a density saliency map DM:

$$DM(x,y) = SM(x,y) * g_\sigma(x,y) \qquad (3)$$

The standard deviation σ is determined in accordance with the accuracy of the eye-tracking apparatus. This filtering also deals with the fact observers stare at a particular areas rather than at a precise point.

2.2. Predicted saliency maps

In previous work, a computational model of the selective visual attention has been designed. This biologically plausible model predicts a saliency map that is in good agreement with the ground truth captured by the eye tracker system [4, 5].
This model is based on a psycho-visual space in which all the early visual data (achromatic and chromatic) are coherently normalized to their own visibility threshold. The most interesting parts of the picture are then identified by simulating the behavior of the cortical cells. These cells have the particularity to suppress the local redundancy of a signal.
Figure 1 shows the experimental and the predicted saliency maps. The brighter areas are the most attractive regions of the picture.

3. SALIENCY-BASED REPURPOSING METHOD

The synoptic of the proposed method shown in figure 2 is described in the following sections.

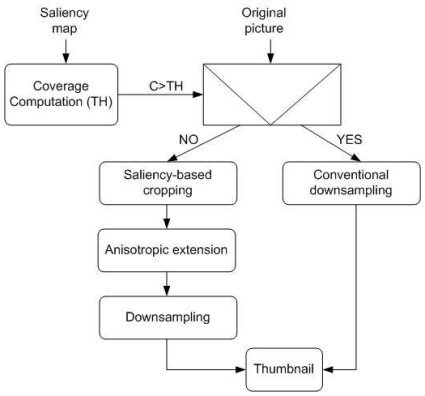

Fig. 2. Flow chart of the proposed method.

3.1. Coverage computation

Coverage has been previously defined by D. Wooding [6] in the following terms: the coverage is a measure of the amount of the original stimulus covered by the fixations. Fixation points are used by the visual system in order to efficiently scan the visual environment. Two consecutive fixations are linked by ballistic saccadic eye movements. The coverage value C is therefore given by the proportion of pixel that have

been fixated. A threshold, called TH, is required in order to decide whether a pixel is fixated or not. This threshold has been experimentally determined.
When the coverage C is important, it means that the picture contains numerous information. In this case, the best thumbnail of the picture is likely to be the picture itself. Conversely, a small coverage value indicates that there is a particular area that attracts the visual attention. Therefore, it is highly probable that a specified rectangle corresponding to this area will be a good thumbnail.

3.2. Saliency-based cropping

In case of small coverage value and given an image I, the goal is to find a subset of I that contains the most relevant areas. These areas are represented by a local maximum in the saliency map. Therefore, the first K maximums of the saliency map are first located. When the k^{th} local maximum is located and memorized, its neighbors (pixel contained in a circle having a radius equal to 1 degree of visual angle) are inhibited in order to determine the spatial location of the $(k+1)^{th}$ local maximum such as a winner-takes-all algorithm. After, only the first m areas ($1 \leq m \leq K$) are considered. The m areas should contained most of the saliency value. All the areas are locally repositioned about their centroid position. This repositioning is necessary to take into account the local distribution of the saliency. The reduced picture corresponds to the rectangle that includes the first m areas. Figure 3 illustrates the determination of the final thumbnail.

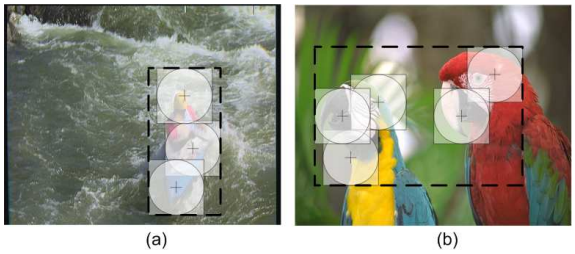

Fig. 3. Saliency-based cropping algorithm for two pictures.

3.3. Anisotropic extension

Thumbnail size is an important parameter for a browsing device. In general, the areas in which the thumbnails have to be displayed are squared. In order to optimize the display, the final reduced picture should have an aspect ratio close to one. An anisotropic extension is then applied on the cropped picture. Two cases are considered: an horizontal (respectively vertical) extension is done when the cropped picture is vertically (horizontally) elongated. Two ratios are considered: $\frac{S_x^R}{S_x}$ and $\frac{S_y^R}{S_y}$. The first and the second ratio are used to detect respectively an horizontally and vertically elongated picture.

Table 1. Experimentally Saliency-based (EST) versus conventional thumbnail (* indicates that the method EST is significantly better than the conventional method ($p < 0.05$, paired t-test)).

Display size	Conventional	EST	No difference
0.6×0.6 inch	32%	47%*	21%
0.9×0.9 inch	36%	43%	21%
1.2×1.2 inch	40%	39%	21%

3.4. Decimation

The thumbnail is finally obtained by down sampling the reduced picture in order to fit the display's size.

4. QUALITATIVE AND QUANTITATIVE ASSESSMENT

Three methods of image retargeting have been tested on twenty five pictures of various contents. First, the conventional approach used by the current device is considered. Thumbnails are obtained by shrinking the original image. The second and the third approach use the method previously described. What distinguishes the two approaches is the saliency map. The first approach, called Experimentally Saliency-based Thumbnail (EST), uses the saliency maps coming from the eye tracking experiments, whereas the second, called Predicted Saliency-based Thumbnail (PST) is based on the saliency map computed by the computational model of the selective visual attention.

4.1. Qualitative assessment

Tables 1 and 2 show the results of the qualitative assessment. Sixteen participants were asked to give theirs preferences (if it exists) between two thumbnails. Three different thumbnail sizes are used 0.6×0.6, 0.9×0.9 and 1.2×1.2 inch corresponding to the size that can be reserved to display a picture (respectively on a mobile phone, a PDA and a PC).
A first test (Table 1) is conducted in order to compare the conventional method (decimation) with the EST method. The second (Table 2) involves the conventional and the PST methods. Both tests indicate that the performances of the saliency-based methods are far superior to the classic thumbnail, whatever the display size. Nevertheless, it is interesting to note that this preference is lessened when the display size increases. It is clear that when the size of the display is large enough, it is not necessary to select a small part. This is preferable in order to keep the picture's context. The PST method outperforms the EST methods due to the fundamental difference that exists between the experimental and the predicted saliency maps; the saliency of the former map is very concentrated whereas

Table 2. Predicted Saliency-based (PST) versus conventional thumbnail (* indicates that the method PST is significantly better than the conventional method ($p < 0.05$, paired t-test)).

Display size	Conventional	PST	No difference
0.6×0.6 inch	23%	52%*	25%
0.9×0.9 inch	25%	53%*	22%
1.2×1.2 inch	31%	49%*	20%

Table 3. Quantitative assessment of EST and PST methods with or without the anisotropic extension (* indicates that the method PST is significantly better than the method PST without anisotropic extension ($p < 0.05$, paired t-test).).

	$\overline{d}$	σ	$MAX_i(d_i)$
IT vs EST without extension	101	29	306
IT vs EST	100	30	250
IT vs PST without extension	121	64	326
IT vs PST	104*	58	326

the latter is smooth, covering numerous parts of the picture, as illustrated on figure 1. It likely due to the fact that participants preferred less cropped images.

4.2. Quantitative assessment

Experimentally Saliency-based and predicted Saliency-based thumbnail, respectively called EST and PST, are compared to ideal thumbnails (IT) that have been chosen in a supervised manner. The comparison is based on the euclidean distance that is computed between the two top-left corners and the two bottom-right corners as illustrated by figure 4. Obviously, a distance equal to zero means that the ideal and the predicted thumbnails are exactly the same. Twenty five pictures are used to evaluate the performances of the EST and PST methods. Table 3 gives the average distance $\overline{d}$, the standard deviation σ and the maximum distance ($MAX_i(d_i)$ where i represents the index of a particular picture). First and foremost, the two methods (EST and PST) provide the same results in average. It means that the proposed computational model simulating the human selective visual attention yields reliable saliency maps for use in this application. Considering now the PST method only, it is noticeable that the anisotropic extension allows to significantly improve the performances. It is important to point out that the standard deviations for PST are double those of EST. It means that the PST method, based on predicted saliency maps, does not succeed to retrieve the visually important areas in all cases, contrary to the EST approach.

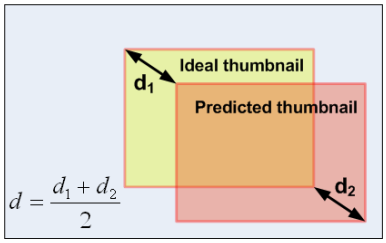

Fig. 4. Ideal and predicted thumbnails of a particular picture are compared by computing the average distance d existing between the two top-left corners and the two bottom-right corners.

5. CONCLUSION

In this paper, a new solution to calculate thumbnail based on visual attention is proposed. A saliency map is deduced from either an eye tracking experiment or from a computational model. Compared to the conventional method, a significant gain, given by a quantitative test, is yielded by the proposed method. This gain is larger for the smaller display size. The new generation of portable media displays should greatly benefit from this new technique. Obviously, the next step of this work would be to consider video sequences.

6. REFERENCES

[1] X. Fan, X. Xie, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, "Visual attention based image browsing on mobile devices," in *ICME 2003*, USA, 2003, vol. 1.

[2] 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," in *ACM Multimedia systems journal*, 2003.

[3] F. Liu and M. Gleicher, "Automatic image retargeting with fisheye-view warping," in *UIST '05: ACM symposium on User interface software and technology*, USA, 2005, pp. 153–162.

[4] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "Performance assessment of a visual attention system entirely based on a human vision modeling," in *Proceedings ICIP-04*, Singapor, 2004, pp. 2327–2330.

[5] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, May 2006.

[6] D. S. Wooding, "Eye movements of large population : Ii. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Research Methods, Instruments and Computers*, vol. 34, no. 4, pp. 509–517, 2002.