Attention-based video reframing: validation using eye-tracking

Christel Chamaret, Olivier Le Meur
Thomson R&D France
1, Av. de Belle-Fontaine - CS 17616 - 35576 Cesson-Sevigne - France
christel.chamaret@thomson.net, olivier.le-meur@thomson.net

Abstract

Watching TV shows on cell phones is starting to become a reality. Nevertheless, there still exist some significant issues due to the small size of cell phone screens. The direct transfer of contents that are not specifically shot for the mobile device will provide indistinguishable objects. An automated way, delivering the best viewing experience is proposed in this paper. This solution significantly improves the visual comfort, by zooming in on the regions of interest. The relevance of this solution rests on its capability to preserve the visually important areas as well as the temporal stability. Eye-tracking experiments are one metric to assess the reframing quality. Involving 16 observers, they show that more than 90% of the visually important regions are kept in the reframed clip.

1. Introduction

Due to the proliferation of cell phones having the capacity of playing video sequence, new viewing experiences on small screen devices are expected. The specific problem of watching video on these new platforms is the screen's size. As most of video contents are not produced with small-screen viewing in mind, the direct transfer of video contents would provide a video on which the main objects of interest may become indistinguishable from the rest of the image. A solution is to focus on the most visually interesting parts of the video and then to repurpose the video content suitably. Today it is generally done manually, repurposing the video content is thus expensive and time consuming. An automated way, delivering the best viewing experience, would be a high economic differentiator. This solution brings two difficulties (figure 1): firstly, the detection (in an automatic manner) of regions of interest or RoIs and secondly, the quality of the reframing.
Many works about automatic reframing deal with still pictures and involve a visual attention model. In [3], they proposed to use the presentation technique, called RSVP (*Rapid Serial Visual Presentation*). This presentation scans sequentially the regions of interest as a human observer will do.
Another approach [8] rests on the use of higher level detectors as face or skin detectors. Those detectors bring cognitive information in order to improve the results stemming from a purely bottom-up visual attention or to drive directly the reframing strategy. The rationale of this approach lies simply on the fact that humans' eye are effortlessly attracted by faces.

Figure 1. From top to bottom: the original frames with cropping window, saliency maps from Visual Attention Model, the cropped frames.

One drawback of previous studies concerns the loss of context when no salient parts of image are discarded. The improvement of previous techniques, proposed by [6], is to use non-linear fisheye-view warp emphazing visually important parts of the picture while shrinking others. Even if the context is preserved, the new picture is distorted. More recently, a seam carving algorithm was used to resize the picture [2]. Results are encouraging but strong reduction of a picture's size may significantly affect the regions of interest. In addition,

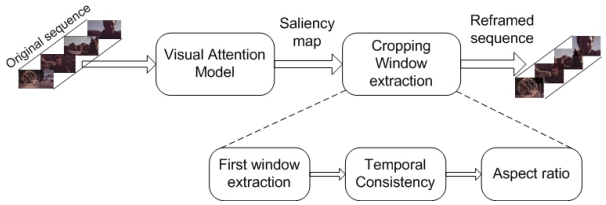

Figure 2. Automatic reframing process.

the image's aspect ratio is not kept since some rows or columns are removed depending on their energy level.
Concerning the video domain, few works have been published. Potential rival techniques belong to the broadcasting world. Three techniques are mainly used in case of format conversion: anamorphic distortion, letter/pillar box and centered cropping. Since those techniques process all frames of the sequence in the same way, there is no potential distortion. Nevertheless, these methods are not driven by the content. The pan-scan format [1] defines a context for conveying metadata containing cropping windows. The proposed application goes further: in addition to the horizontal direction, the cropping window is adaptive vertically and by its size.
This paper deals with the video reframing and its validation. Section 2 provides a description of the proposed automatic solution. Section 3 shows the validation of the proposed algorithm based on eye-tracking user tests.

2. Automatic video reframing

In the same way as previous works on still pictures, the proposed technology rests on both an efficient visual attention model and the cropping window extraction (figure 2). Firstly, a visual attention model computes a saliency map per frame which classifies the frame regions according to their visual interest. Afterwards, a cropping window which encloses the most important parts is deduced from the saliency map.

2.1. Visual Attention Model (VAM)

As our visual environment contains much more information than we are able to perceive at once, it is necessary to select the most important parts of a scene. This task is achieved by our visual attention.
Most of the models provide a topographic saliency sequence which quantitatively predicts how conspicuous every location in the input sequence is. These models are, for most of them, based on the use of the low-level visual features. These models are commonly called bottom-up models. At the opposite top-down models take into account cognitive information as faces, text... or are dedicated to achieve a particular task. In figure 1, the saliency maps highlight the most attractive regions. In this contribution, the model described in [4, 5] is used. This is a purely bottom-up model based on luminance, color and motion information. From an incoming video, a spatio-temporal saliency map is computed.

2.2. From the saliency map to the first cropping window

The goal of this step is to extrapolate a cropping window that encloses the most conspicuous parts of the scene. Based on the results coming from the attention model, a Winner-Take-All algorithm is applied. This algorithm allows the detection of the first N most important locations (having the highest saliency values). When the k^{th} maximum location is selected and memorized, this location as well as its neighborhood is inhibited. Due to the inhibition process, a new saliency peak will dominate and will be selected at the next iteration. As depicted in figure 3, the first N fixation points are detected.

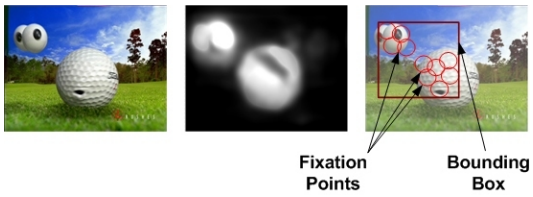

Figure 3. From left to right: original frame, saliency map and the first fixation points enclosed by a cropping window.

2.3. Temporal consistency

The temporal filtering is likely one important issue, because it has to deal with two aspects:

- objects (still or not) or texture may appear along a scene and could lead to a significant change in the distribution of salience. Significant changes of the salience distribution could alter the position and/or the size of the cropping window over time.

- another fundamental idea is to be closed to camera motion. As soon as reframing application behaves as a camera, potential watcher will not make any distinction between a reframed sequence and the original one.

The smoother the temporal cropping process is, the better the visual impact for any users will be. The temporal stabilization acts both on the position and size of the bounding box. Temporal consistency is composed by two sequential steps: a Kalman filter is first applied in order to better predict the current samples. Then, a temporal filtering allows avoiding unlikely samples.
Regarding the position of the bounding box centre and its size, two discrete Kalman filters are used to ensure a good temporal consistency. Rather than working on all of the previous data to provide an estimation (as a Wiener filter will do), the Kalman filter does not require the previous data. It ensures a smooth trajectory and a natural zoom effect of the cropping window. The process is reinitialized after a scene cut in order to avoid the propagation of trajectory regarding the previous scene.
The temporal filtering process is applied on three parameters featuring the bounding box: its center coordinates and its size.

2.4. Aspect ratio: anisotropic extension

The aspect ratio gives the relationship that exits between the width and the height of the original sequence. The first extracted window can not correspond to the final aspect ratio: a priori, there is no relation between the window extracted from saliency map and the user settings. In some way, one side has to be extended. The extension is either on width or on height to reach the targeted aspect ratio. The choice between those two directions is performed according to the aspect ratio: If $\frac{Size_Y^T}{Size_X^T}$ (the target aspect ratio) $> \frac{Size_Y^R}{Size_X^R}$ (the reframed aspect ratio after the temporal consistency step), the extension is horizontal (it will be performed on the width of the current rectangle) else the extension is vertical (it will be performed on the height of the current rectangle) with $Size_X^T$ and $Size_Y^T$ are the frame width and the frame height defined by the user, respectively, and with $Size_X^R$ and $Size_Y^R$ are the frame width and the frame height after the temporal consistency step, respectively. Once the side of extension is defined, there are still several ways to extend the window. Let us assume that the width has to be extended to reach the final aspect ratio. The extension may be entirely transferred to the left side (figure 4 (b)), to the right side (figure 4 (a)) or to both sides in the same proportion (figure 4 (c)). Those solutions are not optimal from a content point of view. Figure 4 illustrates the results provided by the proposed solution: it extends the window side according to the saliency spreading. If there is more saliency in the area of one direction, the cropping window will be extended in that direction proportionally to the saliency. In figure 4 (d), the man at the right-hand side of the ball is enclosed, while he is not in case of Figure 4 (b) and (c).

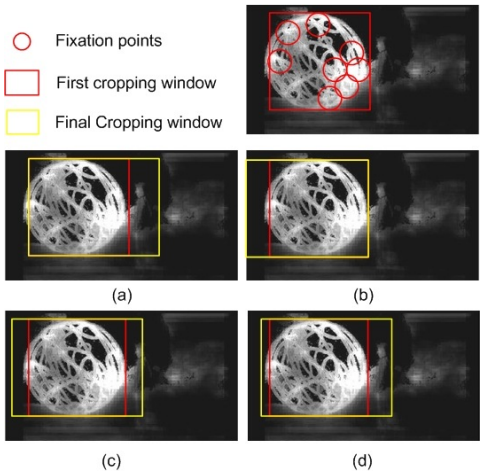

Figure 4. Aspect ratio extension.

3. Eye-tracking validation

It concerns the ability of the reframing to keep most visually important information in the final result.

3.1. Method

Sixteen unpaid subjects participated to the experiments. All observers had normal or corrected to normal visual acuity and normal color perception. All were inexperienced observers (not expert in image or video processing) and naive to the experiment.
Eye movement recording has been performed with a dual-Purkinje eye tracker from Cambridge Research Corporation. Experiment began with a calibration procedure done for each observer. Once the calibration procedure is completed and a stimulus has been loaded, the system is able to track subject's eye movements. Video was processed in real-time to extract the spatial location of the eye position. Both Purkinje reflections are used to calculate the eye's location. The guaranteed sampling frequency is $50Hz$ and the accuracy is about 0.5 degree. Four video sequences have been selected: *Movie*, *Cartoon1*, *Cartoon2* and *Sports*. Each sequence was presented to subjects in a free-viewing task. Experiments were conducted in normalized conditions (ITU-R BT 500-10). The spatial resolution of video sequence is 720×480 with a frequency of $50Hz$ in a progressive mode. They are displayed at a viewing distance of four times the height of the picture ($66cm$). Subjects were instructed to look around the image. The objective is to encourage a visual bottom-up behavior.

3.2. Reframing validation

After the tests, the number of visual fixations that fall into the cropping window is computed. Figure 5 gives this percentage. Whatever the tested clips, more than 90 percents of human fixation points are included within the cropping window.
On the same figure, two others information is given: the minimum percentage as well as the average value of the lowest values (10% of the lowest values). The former is about 20% for the *Movie1* and *Sports* clips and greater than 60% for the other clips. These relatively low values are due to the temporal masking induced by a scene cut [4]. After a scene cut, the spatial coordinates of the visual fixation depend on the content displayed prior the cut. This temporal shifting is due to the inability of visual system to instantaneously adjust to changes. Previous studies demonstrated that the perception is reduced after a brutal changes and can last up to $100ms$ [7]. Therefore, just after a scene cut, the cropping window is well located whereas the position of the human's gaze is still locked on areas corresponding to the content prior the cut.
The latter information is more reliable since the lowest values (10% of the considered data) are averaged. Results are then less noisy by the temporal masking. These values suggest that the loss of visually important areas is well mastered. The worst value is obtained when the *Sports* clip is considered (60%). Indeed, this kind of content contains numerous regions of interest and the consistency in fixation locations is not as high as those obtained by animated sequences or movie clips.
These results indicate that areas where people look at are well selected by the visual attention model as well as enclosed by the cropping window.

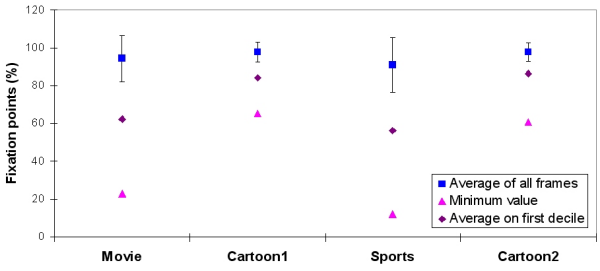

Figure 5. Percentage of human fixation points in cropping window.

4. Conclusion and future work

This paper described a new automatic video reframing process. Most of previous works are dedicated to still pictures. The proposed algorithm is based on a visual attention model which first identifies some regions of interest. The heart of the reframing process consists of the extraction of a cropping window which is both related to the RoI and temporally smoothed in terms of location (center coordinates) and size by means of a strong temporal filtering.
The resulted sequences are difficult to assess: more than a user feeling, it was required to find a quality measurement method. In that way, the eye-tracking apparatus appears to be an interesting tool. The eye-tracking results of some volunteers for the considered sequences were recorded. Finally, more than 90 percents of user fixation points are inside the cropping window whatever the sequence content. Consequently, the automatic video reframing application succeeds in enclosing most of regions where people watch.
Future work mainly consists in improving the VAM. Some cognitive information as face detectors may reinforce the model. Globally, a better understanding of the video content will improve the overall performances of the solution.
The authors would like to thank the University of Nantes and Alexandre Ninassi for eye-tracking tests.

References

[1] Format for pan-scan information. *SMPTE STANDARD, SMPTE 2016-2-2007*, 2007.
[2] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *SIGGRAPH 2007*, volume 26 of *ACM Transactions on Graphics*, 2007.
[3] X. Fan, X. Xie, W. Ma, H. Zhang, and H. Zhou. Visual attention based image browsing on mobile devices. In *ICME 2003*, volume 1, pages 53–56, 2003.
[4] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483–2498, 2007.
[5] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model the bottom-up visual attention. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(5):802–817, 2006.
[6] H. Liu and M. Gleicher. Automatic image retargeting with fisheye-view warping. *18th annual ACM Symposium on User interface software and technology*, pages 153–162, 2005.
[7] A. J. Seyler and Z. Budrikis. Details perception after scene changes in television image presentations. *IEEE Trans. Inform. Theory*, 11(1):31–43, 1965.
[8] B. Suh, H. Ling, B. Bederson, and D. Jacobs. Automatic thumbnail cropping and its effectiveness. *UIST '03*, pages 95–104, 2003.