Visual Attention Driven Image to Video Transmoding

Joel Baltazar, Pedro Pinho, Fernando Pereira

Instituto Superior Técnico - Instituto de Telecomunicações
IST, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
fp@lx.it.pt

Abstract. Nowadays, the heterogeneity of networks, terminals, and users is growing. At the same time, the availability and usage of multimedia content is increasing, which has raised the relevance of content adaptation technologies able to fulfill the needs associated to all usage conditions. For example, mobile displays tend to be too small to allow one to see all the details of an image. In this report, a solution for this problem is proposed: an automatic adaptation system, that uses visual attention models, creates a video clip that browses through the image displaying its regions of interest in detail.

This paper describes the architecture developed for the adaptation system, the processing solutions and also the principles and reasoning behind the algorithms that have been developed and implemented to achieve the objective of this work. In order to evaluate the performance of the adaptation system, a user study has been conducted. The results of the study are encouraging, since they indicate that users consider the quality of the experience provided by the video clips to be better than the still image experience.

Keywords - Transmoding, Video Adaptation, Image Browsing, Visual Attention, Regions of Interest

1. INTRODUCTION AND OBJECTIVES

Networks, terminals and users are becoming increasingly heterogeneous. In this context, the growing availability and usage of multimedia content have been raising the relevance of content adaptation technologies able to fulfill the needs associated to all usage conditions, without multiplying the number of versions available for the same piece of content. This means that adaptation tools are becoming increasingly important to provide different presentations of the same information that suit different usage conditions. Furthermore, the importance of the user and not the terminal as the final point in the multimedia consumption chain is becoming clear.

Nowadays people share many of their important moments with others using visual content such as photographs, that they can easily capture on their mobile devices anywhere, at anytime. Therefore, images are very important in mobile multimedia applications. However, mobile devices have several limitations, notably regarding computational resources, memory, bandwidth, and display size. While technological advances will solve some of these limitations, the display size will continue to be a major constraint on small mobile devices such as cell-phones and handheld PC's. Currently, the predominant methods for viewing large images on small devices are down-sampling or manual browsing by zooming and scrolling. Image down-sampling results in significant information loss, due to excessive resolution reduction. Manual browsing can avoid information loss but is often time-consuming for the users to catch the most crucial information in an image. In [1] an adaptation tool allowing the automatic browsing of large pictures on mobile devices is proposed by transforming the image into a simple video sequence composed of pan and zoom movements which are able to automate the scrolling and navigation of a large picture. Similar solutions are proposed in [2] and [3].

In this paper, an adaptation system (called *Image2Video*) whose major objective is to maximize the user experience when consuming an image in a device with a small size display is proposed. The processing algorithms developed to reach this purpose imply determining the regions of interest (ROIs) in an image based on knowledge of the human visual attention mechanisms, and generating a video sequence that displays those regions according to certain user preferences, while taking into consideration the limitations of the display's size, as shown in Figure 1. User preferences refer to the video display modes the user can choose for the visualization of the adapted video sequence, e.g. the duration of the video. The created video is intended to provide a final, better user experience, compared to the down-sampled still image or the manual scrolling alternatives.

2. Proposed Image2Video System Architecture

The developed adaptation system is inspired on the knowledge of the human visual system (HVS) attention mechanisms to determine ROIs in the image, and it uses a multi-stage architecture to perform all the necessary tasks to transform the original image into a (more interesting) video clip.

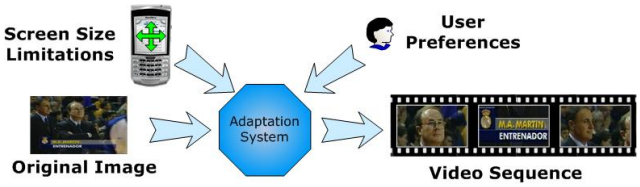


Figure 1 - Image2Video adaptation system

The multi-stage architecture proposed for the adaptation system is presented in Figure 2. It was conceived to produce a high-level description of the most interesting contents in the image and then combine that description with the user preferences and terminal device limitation descriptions to perform the image-to-video transmoding.

Transmoding refers to all media adaptation processes where content in a certain modality is transformed into content in another modality, e.g. video to images, text to speech. The proposed architecture includes four stages, with the first two being responsible for the determination of a map that identifies all the ROIs of the image, and the remaining two responsible for the generation of the video that displays the image following a path that conveniently links the various ROIs to each other.

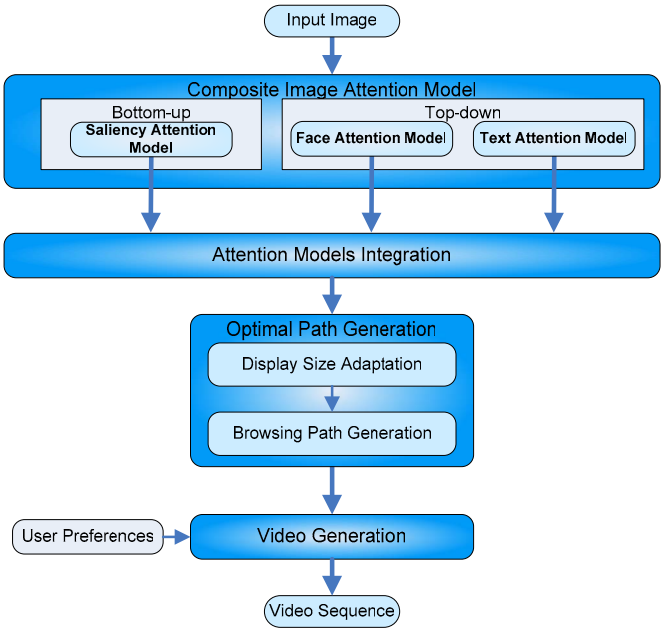


Figure 2 - Image2Video system main architecture

The main objectives of the four main architectural stages are presented in the following sections.

2.1 Composite Image Attention Model

The HVS attention mechanism is able to select ROIs in the visual scene for additional analysis; the ROIs selection is guided by bottom-up and top-down approaches. Using the knowledge of the human visual attention mechanisms, a composite image attention model has been developed to detect ROIs and provide a measure of the relative importance of each one.

Based on the work developed by Chen et al. [2], which proposes a method for adapting images based on user attention, the visual attention models provide a set of attention objects (AOs):

$$\{AO_i\} = \{(ROI_i, AV_i)\}, \qquad 1 \le i \le N$$

Frequently, an AO represents an object with semantic value, such as a face, a line of text or a car, meaning that it carries information that can catch the user's attention. Therefore the ith attention object within the image, AOi, has two attributes: the ROIi, which is the region of the image that contains the AOi; and the attention value (AVi), which represents an estimate of the user's attention on the AO. The basis for the AV is that different AOs carry different amounts of information, so it is necessary to quantify the relative importance of each one.

The model integrates three elementary visual attention models:

- **Saliency Attention Model -** The objective of this model is to identify ROIs without specific semantic value associated objects, i.e. regions with different statistical properties from the neighboring regions are considered ROIs [4].
- **Face Attention Model** - The objective of this model is to identify ROIs that contain faces. The detection of faces is a task performed daily by humans since they are one of their most distinctive characteristics, providing an easy way to identify someone. Therefore faces are one of the semantic objects present in an image that are more likely to captivate human's attention [5].
- **Text Attention Model** - The objective of this model is to identify ROIs that contain text. People spend a lot of their time reading, may it be newspapers, e-mails, SMS, etc. Text is a rich font of information, many times enhancing the message that an image transmits. Therefore text is the kind of semantic object that attracts viewer's attention.

2.2 Attention Models Integration

The attention models integration stage is responsible for integrating all the identified types of AOs into a unique image attention map using pre-defined criteria to solve the cases where spatial overlapping exists between them. The criteria used to solve the three considered overlapping cases, Face-Text, Face-Saliency and Text-Saliency, are now presented:

- **Face-Text integration** - The process to solve the cases where the bounding boxes of text and face ROIs overlap states that they should always remain independent. Face and text AOs have completely different semantic values, and if overlapping exists it is due to imperfections in the definitions of their bounding boxes.
- **Face-Saliency integration** - When face and saliency ROIs overlap, it is necessary to determine if they represent the same object or not, and for this a criterion has been developed. The criterion, expressed by Eq. (2.1), states that only when the face ROI contains a big part of the saliency ROI, they are likely to represent the same AO: a face. Otherwise the two ROIs remain independent.

$$\frac{area(ROI_{face} \cap ROI_{saliency})}{area(ROI_{saliency})} \geq 0.25 \quad (2.1)$$

- **Text-Saliency integration** - In this case, the developed criterion, expressed by Eq. (2.2), states that only when the text ROI contains a big part of the saliency ROI, it is likely they represent the same ROI: text. Otherwise the two ROIs remain independent.

$$\frac{area(ROI_{text} \cap ROI_{saliency})}{area(ROI_{saliency})} \geq 0.25 \quad (2.2)$$

After integrating all the AOs by solving the overlapping cases, it is necessary to distinguish the importance that each AO has according to its type: saliency, face or text. Faces are considered to be very important for humans; this means faces are the kind of object that a human will look for first in an image. Text can provide a lot of information, delivering or enhancing the message that an image is intended to transmit; therefore it is considered the second most important type of AOs. Saliency AOs are considered the least important because nothing is known regarding their semantic value, they can be any kind of object.

Therefore to calculate the final AV of each AO, Equation (2.3) is used, where *Wm* is the weight corresponding to the type of AO, and $m \in \{saliency, face, text\}$. Exhaustive experiments were performed to obtain the following weight values: $W_{Saliency} = 0.2$, $W_{Text} = 0.35$ and $W_{Face} = 0.45$.

$$AV_{final} = AV \times W_m \quad (2.3)$$

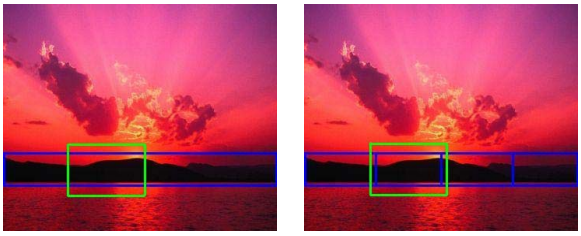
AOs that have a relative small AV are considered to provide little information and therefore AOs that don't fulfill Eq. (2.4) are eliminated.

$$\frac{AV(AO_i)}{\max AV(AO_j)} \geq 0.10 \qquad j = 1,..,N \qquad (2.4)$$

2.3 Optimal Path Generation

This stage is responsible for generating the path used to display with video the whole image, i.e. the path that transforms the image into video. Two mechanisms are used for this:

- **Display Size Adaptation** - The video sequence is created so that AOs are displayed with their maximum quality, i.e. the AOs are displayed with their original spatial resolution. The objective of this mechanism is to optimize the information presented on the screen at each moment. To do so, the developed algorithm splits or groups AOs to form attention groups (AGs), which have a dimension equal to or smaller than the display size. The display size adaptation process is performed using two methods: first Split Processing and afterwards Group Processing:
    1. ***Split Processing*** **-** It is usual that the spatial resolution of the AO is bigger than the spatial resolution of the image display in which case it is necessary to spatially divide the AO in smaller parts that fit the display size. Face AOs are never divided since they must be visualized as a whole to have semantic meaning. Figure 3 (a) presents an example of a ROI whose horizontal dimension exceeds the horizontal size of the display (green rectangle), and therefore is divided into four AGs that fit into the display size (see Figure 3 (b)).



(a) (b)

Figure 3 - Example of AG split processing

2. ***Group Processing*** - Since some AGs are very small compared to the display size they can be grouped with others, when possible, forming an AG which provides maximum information to the user on the display. Figure 4 shows an example where grouping is possible.

(a) (b)

Figure 4 - Example of Group Processing

- **Browsing Path Generation** - This mechanism determines the order by which AGs will be displayed, and therefore establishes the path that will be used to display with video the whole image. AGs are shown in detail, following the order of their AV, i.e. the AG with the highest AV is the first to be displayed. However, in some cases the displaying order can be changed. Changing the order by which AGs are displayed can save displaying time, and also avoid traveling back and forward in the image, which can be unpleasant for the user. Figure 5 presents an example of the spatial distribution of three AGs with $AV(AG_i) > AV(AG_j) > AV(AG_k)$, which is used to explain the developed criteria to decide in which cases the displaying order of the AGs should be changed:

  1. ***AGs Distance Criterion*** - This criterion states that if the traveled distance for the normal order of the AGs, and the distance traveled by attending another AG_k is similar, the displaying order of the AGs could be changed. The criterion is checked if Equation (2.5) is fulfilled, i.e. if the distance ratio is equal to or bigger than 95%.

$$\frac{dist(AG_i, AG_j)}{dist(AG_i, AG_k)+dist(AG_k, AG_j)} \geq 0.95 \quad (2.5)$$

  2. ***AGs Attention Value Criterion*** - This criterion states that if the AVs of the second and third most important AGs, AG_k and AG_j, are similar then the displaying order could be changed. This criterion is fulfilled if Equation (2.6) is verified, i.e. if the AVs ratio is equal to or bigger than 90%.

$$\frac{AV(AG_k)}{AV(AG_j)} \geq 0.90 \quad (2.6)$$

When both the criteria are fulfilled, the order by which AG_i, AG_j and AG_k are displayed becomes as in Figure 5.

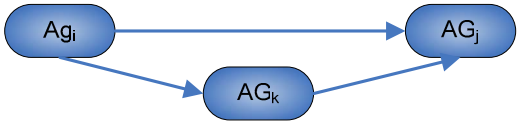


Figure 5 - Example of spatial distribution of AGs

Figure 6 shows an example of the optimal path to attend the AGs present in the image. The numbers inside the bounding boxes represent the order by which AGs are displayed.

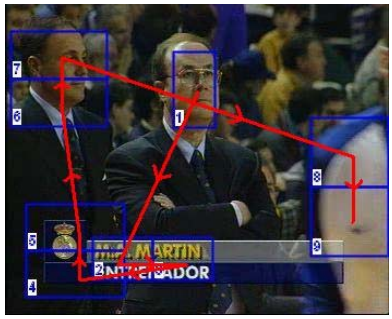


Figure 6 - Example of browsing path

2.4 Video Generation

The last stage of the adaptation system is responsible for creating a video sequence based on the previously calculated browsing path, a set of directing rules and user preferences. As different users have different needs and different preferences, one of the three different display

modes for the visualization of the adapted video sequence can be chosen:

1. **Normal:** All the AGs are presented, without any restriction.
2. **Time Based (TB):** The user chooses to see an adapted video sequence with a maximum time limit; therefore the video sequence will only show the most important AGs within the time limit.
3. **Amount of Information Based (AIB):** The user chooses to see an adapted video sequence with a determined percentage of information, i.e. the video sequence will only show the most important AGs corresponding to the chosen percentage of information.

3 User Evaluation Study

The purpose of the user study that has been conducted is not only to evaluate the global performance of the developed adaptation system, but also to assess the performance of the developed algorithms. As there is no objective measure to evaluate the performance of the developed adaptation system, the user study provides a subjective evaluation of the results provided by the Image2Video application, having three main objectives:

1. Evaluate how good the adapted video clip experience is regarding the still image experience (using a downsampled image with a resolution corresponding to the screen resolution) in display size constrained devices, to determine the impact of the application on the user.
2. Evaluate if all the important ROIs in the image are focused in the video clip, and therefore determine the performance of the composite image attention model and attention models integration stages of the adaptation system.
3. Compare the display order of the focused ROIs in the video clip with the order the user would focus on the same ROIs, allowing to determine the adequateness of the calculated AVs and the performance of the optimal browsing path algorithm.

A set of 8 images with a resolution of 352×288 pixels was selected. The images are divided into four classes:

- **Saliency class***:* The images in this class don't contain human faces or text.
- **Face class:** The images in this class contain human faces and no text.
- **Text class:** The images in this class contain text and no human faces.
- **Mix class:** The images in this class contain both human faces and text.

Based on these 8 images, the adapted video clips where produced with a resolution of 110×90 pixels (corresponding to the display resolution), to simulate viewing the image and the video clip in a display size constrained device. Using the developed application interface to show the original image and the video clip in a small size display, a group of 15 volunteers were invited to give their subjective judgments at the following three questions:

Question 1: How good is the video experience regarding the still image experience ?

a) Very bad b) Bad c) Reasonable d) Good e) Very good

Question 2: Are all the interesting regions of the image focused on the video?

a) None b) Some c) Almost all d) All

Question 3: How well does the focused regions order reflect their real relative importance?

a) Very bad b) Bad c) Reasonable d) Well e) Very well.

Based on the answers provided by the volunteers involved in this study, Table 6.1, Table 6.2 and Table 6.3 contain the statistical results for all three questions. Regarding Question 1, the average results show that 39% and 33% of the inquired consider the video experience compared to the still image experience, good and very good, respectively. These results allow concluding that the majority of the users prefer the adapted video clip instead of the still image. Regarding Question 2, the average results show that 59% of the inquired consider that all of the interesting regions of the image are focused in the video. The average results for Question 3 show that the 41% and 33% of the inquired consider that the ordering of the focused regions reflects their real relative importance, well and very well, respectively.

Based on the evaluation study results, it is possible to conclude that the developed Image2Video application achieves its main objective, i.e., the quality of the experience provided by the adapted video clips created with the proposed application is better than the

experience provided by the downsampled still image experience.

Table 1 – Evaluation results for Question 1

Image Class	a)	b)	c)	d)	e)
Saliency	0%	3%	33%	40%	24%
Face	0%	7%	17%	33%	43%
Text	0%	6%	20%	47%	27%
Mix	0%	3%	23%	34%	40%
Average	0%	5%	23%	39%	33%

Table 2 – Evaluation results for Question 2

Image Class	a)	b)	c)	d)
Saliency	0%	13%	50%	37%
Face	0%	3%	23%	74%
Text	0%	3%	33%	64%
Mix	0%	7%	30%	63%
Average	0%	7%	34%	59%

Table 3 – Evaluation results for Question 3

Image Class	a)	b)	c)	d)	e)
Total Saliency	0%	3%	30%	37%	30%
Face	0%	3%	20%	57%	20%
Text	0%	3%	17%	40%	40%
Mix	0%	3%	27%	30%	40%
Average	0%	3%	23%	41%	33%

4 Conclusions

In this paper, an adaptation system has been presented, which has the objective of maximizing the user experience when consuming an image in a device with a small size display such as the very popular mobile phones. The adaptation is performed using an innovative method: transforming images into video driven by visual attention, targeting a final better user experience.

The data fusion performed by the Attention Models Integration module and the Browsing Path Generation algorithm represent the major innovative contributions of this work: the first provides a unique attention map with all the ROIs of the image, and the second determines the optimal path to browse through the ROIs. Another important contribution is the definition of a modular architecture to transform images into video, making it highly scalable: this means that either a better saliency attention model or new specific object detectors could be added to improve the detection of ROIs in the image.

The results of the user study are encouraging, since they indicate that users consider the quality of the experience provided by the video clips to be largely better than the usual still image experience.

References

[1] H. Liu, X. Xie, W.Y. Ma, H.J. Zhang, “Automatic browsing of large pictures on mobile devices”, ACM Multimedia’2003, Berkeley, CA, USA, November 2003.

[2] L. Chen, X. Xie, X. Fan, W. Ma, H. Zhang, H. Zhou, “A visual attention model for adapting images on small displays”, ACM Multimedia Systems Journal, Vol.9, No.4, pp. 353-364, 2003.

[3] X. Fan, X. Xie, W. Ma, H.J. Zhang, H. Zhou, “Visual attention based image browsing on mobile devices”, ICME’2003, Baltimore, USA, July 2003.

[4] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259, 1998.

[5] J. Ascenso, P. L. Correia, F. Pereira; “A face detection solution integrating automatic and user assisted tools”, Portuguese Conf. on Pattern Recognition, Porto, Portugal , Vol. 1 , pp. 109 - 116 , May 2000.

[6] D. Palma, J. Ascenso, F. Pereira, “Automatic text extraction in digital video based on motion analysis”, Int. Conf. on Image Analysis and Recognition (ICIAR’2004), Porto - Portugal, September 2004.