Integrating Low-Level and Semantic Visual Cues for Improved Image-to-Video Experiences

Pedro Pinho, Joel Baltazar, 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. This paper presents an innovative method to integrate low-level and semantic visual cues into a unique visual attention map that represents the most interesting contents of an image, allowing the creation of a video sequence that browses through the image displaying its regions of interest in detail. The architecture of the developed adaptation system, the processing solutions and also the principles and reasoning behind the algorithms that have been developed and implemented are presented in this paper. Special emphasis is given to the integration of low-level and semantic visual cues for the maximization of the image to video adapted experience.

1. Introduction and Objectives

With the explosion of digital image and video technology, it is nowadays largely felt and recognized that we live in a visual age where multimedia information is present everywhere, anytime. Television is no longer the only way to access multimedia content, since recent technological developments have opened new frontiers to the consumption of multimedia content, notably following an everywhere, at anytime paradigm. This resulted in a growing heterogeneity of networks, terminals and users, and an increase in the availability and usage of multimedia content. The diversity of users, content, networks and terminals has prompted the development of adaptation tools to provide different presentations of the same information that suit different usage conditions based on the principle that every single user should get the best possible experience for the content desired under the relevant consumption conditions.

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 that allows 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, a new method to integrate low-level and semantic visual cues to create an image attention map is proposed. The processing algorithms developed to achieve this objective are based on the human visual attention mechanisms. Building on this method, an adaptation system (called Image2Video) has been developed, which generates a video sequence that displays an image's regions of interest in detail according to certain user preferences, e.g. video duration, while taking into consideration the limitations of the display's size, as shown in Fig. 1.

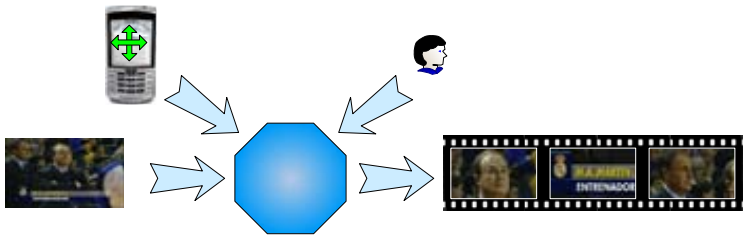


Fig. 1. Image2Video adaptation system

2. Proposed Image2Video System Architecture

The developed adaptation system is inspired on the knowledge of the human visual system (HVS) attention mechanisms to determine regions of interest (ROIs) in the image. Moreover it uses a multi-stage architecture to perform all the necessary tasks to transform the original image into a (more interesting user experience) video clip. The proposed architecture for the adaptation system, presented in Fig. 2, includes four stages which 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. Building on 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):

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

Frequently, an AO represents an object with semantic value, such as a face, a line of text or an interesting object, 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 AOi. 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.

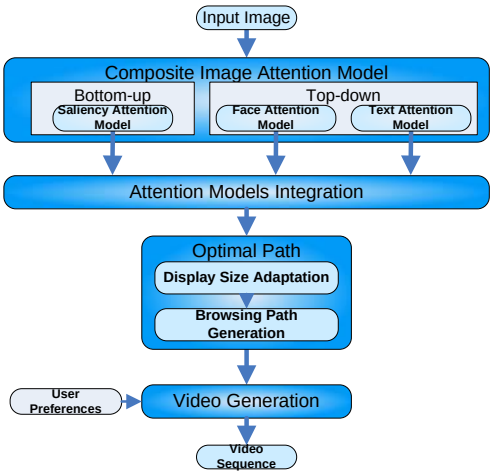
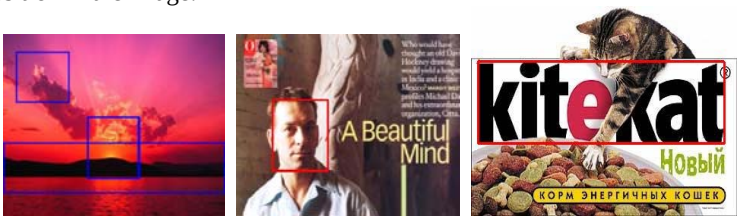


Fig. 2. Image2Video system main architecture

The composite visual attention model integrates three elementary visual attention models (see Fig.3 for example results):

- **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 identified as ROIs. In this model, the AV ($AV_{saliency}$) is determined based on the time difference between the identification of 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 capture human's attention. In this model, the AV (AV_{face}) is based on the AO's area and position in the image [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 a kind of semantic object that attracts viewer [6]. In this model, the AV (AV_{text}) is based on the AO's area and position in the image.



(a) (b) (c)

Fig. 3. Example of regions of interest detection results: a) saliency; b) face; c) text

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 considered to be due to imperfections in their bounding boxes. Therefore it has been decided to trust the ROIs identification provided by the text and face detectors when this type of overlapping exists, i.e. when face and text ROIs overlap they remain independent.
- **Face-saliency integration** - When face and saliency ROIs overlap, it is necessary to determine if they represent the same object or not. The criterion for this decision, expressed by Eq. 2, 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.

$$\frac{area(ROI_{face} \cap ROI_{saliency})}{area(ROI_{saliency})} \geq 0.25 \quad (2)$$

Fig. 4 (a) presents an example where the bigger bounding box representing a face ROI is overlapped by a smaller saliency ROI. Since the criterion defined by Eq. 2 is fulfilled, the saliency ROI is eliminated as shown in Fig. 4 (b).

Fig. 5 shows the structure of the algorithm proposed to solve the cases where face and saliency ROIs overlap. The algorithm tests all the possible combinations of ROIs. For each pair of ROIs, the algorithm verifies if it is a face-saliency pair. When a face-saliency pair exists, the overlap area is calculated. If the face-saliency overlap criterion is fulfilled, the saliency ROI is eliminated, and therefore the number of ROIs decreases. The algorithm then proceeds to the next iteration, i.e. evaluates the next

ROIs pair. The algorithm stops when all possible combinations of face-saliency ROIs pairs have been tested.

(a) Face-saliency ROIs overlap (b) Face ROI prevails, saliency ROI is eliminated

Fig. 4. Example of face-saliency integration

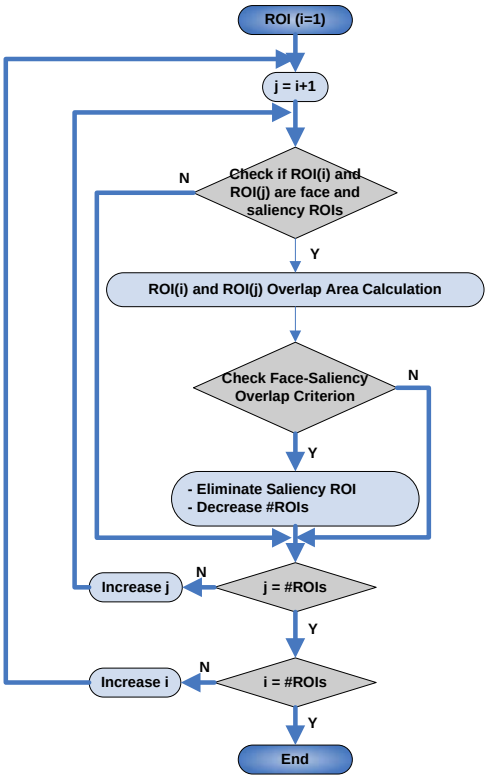


Fig. 5. Structure of the face-saliency ROIs integration algorithm

- **Text-saliency integration -** In this case the proposed decision criterion, expressed by Eq. 3, 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.

$$\frac{area(ROI_{text} \cap ROI_{saliency})}{area(ROI_{saliency})} \geq 0.25 \quad \textbf{(3)}$$

Fig. 6 (a) presents an example where the bigger bounding box representing a text ROI is overlapped by a smaller saliency ROI. Since the criterion defined by Eq. 3 is fulfilled, the saliency ROI is eliminated as shown in Fig. 6 (b). The structure of the algorithm developed to detect and solve the cases where overlapping exists between text and saliency ROIs is similar to the one proposed for face-saliency overlapping.

(a) Text-saliency overlap (b) Text ROI prevails, saliency ROI is eliminated

Fig. 6. Example of text-saliency integration

After integrating all the AOs by solving the overlapping cases, it is necessary to evaluate the importance of each AO, depending on 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 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 AO. 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, Eq. 4 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 $W_{face} = 0.45$.

$$AV_{final} = AV_m \times W_m \quad \textbf{(4)}$$

AOs that have a relative small AV are considered to provide little information and therefore AOs that don't fulfill Eq. 5 are eliminated.

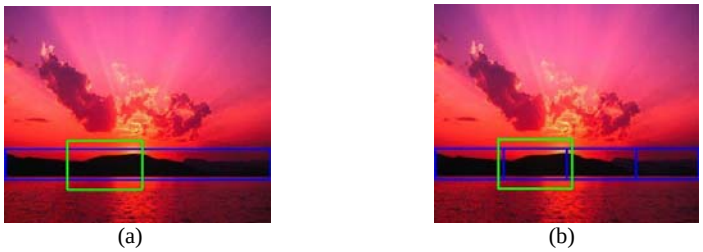
$$\frac{AV_{final}(AO_i)}{\max AV_{final}(AO_j)} \geq 0.10 \qquad j = 1,..,N \quad \textbf{(5)}$$

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. As the AOs, the AGs have two attributes: the ROIi, which is the region of the image that contains the AGi; and the attention value (AVi), which represents an estimate of the user's attention on the AG. An AG can contain part of an AO, one or even more AOs. When an AO is split into two or more parts, that can fit into the display size, they constitute a new AG called twin. When one or more AOs are grouped, they constitute a new AG. If an AO is neither split nor grouped, it also constitutes an AG. The display size adaptation process is performed using two methods:

  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 into smaller parts that fit the display size. Face AOs are never divided since they must be visualized as a whole to have semantic meaning. Text AOs can be divided since the resulting AGs will be displayed in a sequential manner that will allow the user to read the text. Figure 7 (a) presents an example of a ROI whose horizontal dimension exceeds the horizontal size of the display (shown with the green rectangle), and therefore is divided into four AGs that fit into the display size (see Figure 7 (b)).



(a) (b)

Fig. 7. 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. When AGs are grouped, the ROI of the new AG is represented by the smallest bounding box that can contain both AGs, and its AV inherits the highest AV of the grouped AGs. Fig. 8 shows an example where grouping is possible.

(a) (b)

Fig. 8. Example of AG 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. Fig. 9 shows two examples of the optimal path to attend the AGs present in the images. The numbers inside the bounding boxes represent the order by which AGs are displayed.

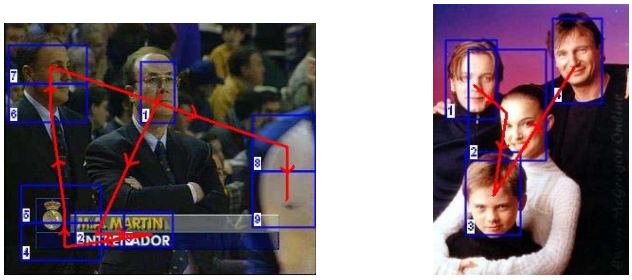


Fig. 9. Examples of browsing paths

2.4 Video Generation

This stage of the Image2Video system is responsible for generating the frames that compose the video sequence, which displays the AGs of the image based on the previously calculated optimal path, a set of directing rules and user preferences. It is important to generate video sequences with smooth motion, so that the video experience is pleasant for the user, in this case also more pleasant than the simple image experience. In order to achieve this target, it is necessary to define the key-frames involved in the motion sequences to be produced.

Typically, in film making, there are three types of video shots: full, medium and close-up. All three have the same spatial resolution, but represent contents of the image with different dimensions, i.e. an object which is small in a full shot can be viewed in detail in a close-up shot where it has a dimension close to the display's size.

Based on the three types of video shots, and in order to generate the video sequence, three types of key-frames have been used [7]:

- **Full-frame (FF) -** Frames that present the entire content of the image; when the aspect ratio of the image and the display are different, the aspect ratio of the image is preserved.
- **Medium-frame (MF) -** Frames used to contextualize the AGs within the image; therefore a window containing 25% of the image, with the aspect ratio of the display, is used. The tests carried out indicated that this window dimension is adequate to perceive the context of AGs. Their spatial dimensions cannot be smaller than those of the display. The MF of an AG is centered on it, therefore contextualizing the AG in the image.
- **Close-up (CF) -** Frames with the dimension of the display size, used to display the AGs in detail.

Fig. 10 presents an example of the content presented by the three types of key-frames: FF (pink), MF (blue) and CF (red).

Fig. 10. Example of the three key-frame types used for video generation

In order to generate the video, it is necessary to create a sequence of key-frames that displays the AGs of the image according to the calculated optimal path. The video sequence must be as smooth as possible in the sense that transitions from key-frame to key-frame should take the time necessary for the user to perceive all the information available. Therefore, experiments were carried out that allowed establishing the default minimal perceptible times (MPT) for the different types of AGs. 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:

- **Normal -** All the AGs are presented, without any restriction.
- **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.
- **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.

4. Image2Video Application Interface

The application interface for the Image2Video system has a main window with three types of boxes as shown in Fig. 11:

- **Input image** - As the name suggests this box situated in the top-left corner of the main window displays the input image chosen by the user.
- **Intermediate results** - The three boxes situated in the lower part of the main window are used to display the intermediate results of the algorithms.
- **Final result** - The box situated in the top-right corner displays the final result of the stages that compose the architecture of the adaptation system.

The main window also contains four action buttons: three buttons to execute the algorithms regarding the stages that compose the architecture of the adaptation system, and one button to open the video player. The *Select Intermediate Results* box allows the user to choose what is displayed in the intermediate results image boxes.

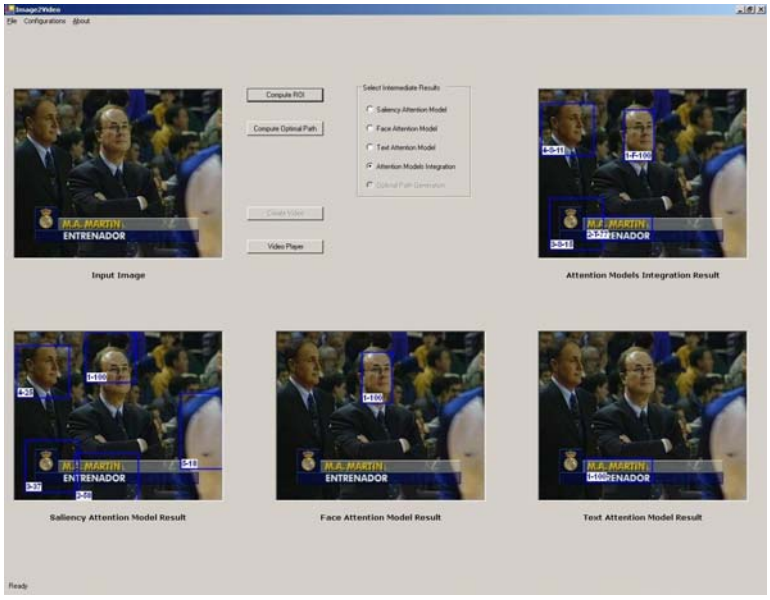


Fig. 11. Main window of the developed Image2Video application interface

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, by using 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.

The images for this experiment were divided into four classes: saliency, face, text and a mix of the previous three classes. For the evaluation, a set of 8 images with a resolution of 352×288 pixels was selected. Based on these 8 images, the adapted video clips where produced with a resolution of 110×90 pixels (corresponding to the test display resolution), to simulate viewing the image and the video clip in a display size constrained device. A group of 15 volunteers was invited to give their subjective judgments regarding the three questions above.

Table 1, Table 2 and Table 3 contain the results for all three questions. Regarding Question 1, the average results show that 39% and 33% of the inquired considered 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 video clip instead of the still image. Regarding Question 2, the average results show that 59% of the inquired considered 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.

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)
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%

5. Conclusions

In this paper, an adaptation system has been developed, with a major objective: maximize 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.

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 down-sampled still image.

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.

[7] Xian-Sheng HUA, Lie LU, Hong-Jiang ZHANG, "Automatically Converting Photographic Series into Video", ACM Multimedia 2004, New York, USA, October 2004.