

# Video summarization using a visual attention model

Sophie Marat, Mickaël Guironnet, Denis Pellerin

► **To cite this version:**

Sophie Marat, Mickaël Guironnet, Denis Pellerin. Video summarization using a visual attention model. 15th European Signal Processing Conference (EUSIPCO-2007), Sep 2007, Poznan, Poland. 5 p, 2007, 3-7 September 2007, Poznan, Poland. <hal-00164604>

**HAL Id: hal-00164604**

**<https://hal.archives-ouvertes.fr/hal-00164604>**

Submitted on 21 Jul 2007

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# VIDEO SUMMARIZATION USING A VISUAL ATTENTION MODEL

*Sophie MARAT, Mick  l GUIRONNET, and Denis PELLERIN*

Grenoble Images Parole Signal Automatique (GIPSA-lab) (ex LIS)  
46 Avenue F  lix Viallet, 38031 Grenoble, France  
phone: +(33) 4.76.57.43.55, fax: +(33) 4.76.57.47.90, sophie.marat@lis.inpg.fr  
www.gipsa-lab.inpg.fr

## ABSTRACT

*This paper presents a method of video summarization based on a visual attention model. The visual attention model is a bottom-up one composed of two parallel ways. A static way, biologically inspired, which highlights salient objects. A dynamic way which gives information about moving objects. A three steps summary method is then presented. The first step is the choice between the two kinds (static and dynamic) of saliency maps given by the attention model. The second step is the selection of keyframes. An "attention variation curve" which highlights changes on frames content during the video is introduced. Keyframes are selected on this variation attention curve. To evaluate the summary a reference summary is built and a comparison method is proposed. The results provide a quantitative analysis and show the efficiency of the video summarization method.*

## 1. INTRODUCTION

With the development of numerical technologies and Internet the volume of videos has increased spectacularly. Therefore it becomes more and more difficult to retrieve interesting videos or excerpts of video and applications such as video summarization, classification or video browsing are now required. Video summarization goal is to make navigation inside video bases and extraction of important events easier. Two kinds of video summary exist: static ones or dynamic ones, called video skimming. Static video summaries reduce video to few representative frames called keyframes. They are presented like a storyboard, while video skimming is presented like movie previews. In this paper, we consider static video summary which need less storage space.

Most of the video summaries are based on low level features (colour, texture,...) [1][2] which are not representative of the video semantic content. We describe a method of automatic static video summarization using a higher level feature given a visual attention model. This attention model generates saliency maps, which point out the areas of frames containing more information and attractive for human gaze.

These saliency maps are used to detect changes on frames of the video which make it possible to select keyframes.

The rest of the paper is organised as follows. The attention model used for summarization is presented in section 2. The summarization method is exposed in section 3. In section 4, a summary evaluation method is described. The results are discussed in section 5.

## 2. ATTENTION MODEL

The most famous visual attention model is the one proposed by Itti and Koch [3]. It is a bottom-up (stimulus driven) model based on features like colour, intensity and orientation; it returns a saliency map per frame. A saliency map is a grey scale frame where bright areas that correspond to the regions which attract human gaze. This static model which considers frames one by one, has been improved recently by integrating motion [4].

Ma et al [5]. have also proposed an user attention model. This model uses a lot of features like static saliency, motion saliency, camera motion, face recognition, aural saliency, keywords, keytopics,... An attention modelling is built on visual attention, aural attention and linguistic attention. This user attention modelling is used for video skimming and a static video can be deduced. This attention model is too complex for static summary only. All the aural features are completely useless for static summary.

The attention model we use is another bottom-up model [6]. It takes motion into account and is more simpler than the one proposed by Ma. It is built on two parallel ways (figure 1). A static way used to extract textured and contrasted regions of frame. This biologically inspired way gathers a retinal filter, a bank of Gabor filters and interactions between filters answers. A dynamic way is used to detect moving objects by an estimation and a compensation of camera motion. This model gives, after temporal filtering and normalization, a saliency map for each way.

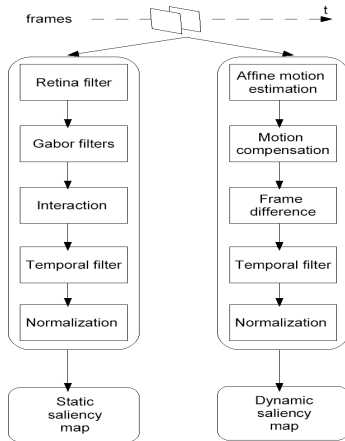


Figure 1 – Principle of attention model.

### 3. VIDEO SUMMARIZATION METHOD

A video summary from attention model has been proposed by Ma et al. [5]. It consists in converting the succession of saliency maps into an “attention curve” and selecting keyframes at the maxima of this attention curve. The attention curve is obtained by replacing each saliency maps by the value of the average of its grey levels. The maxima of this curve correspond to the more salient and then contrasted and textured frames. The drawback of this method is that it keeps temporally nearby frames that are likely to have too similar content.

#### 3.1 Method principle

In this paper we propose a video summarization method based on the attention model presented in figure 1. It is composed of three steps: the choice of the kind of saliency maps to use (static or dynamic), then the keyframes selection, and finally an elimination of redundant frames.

##### 3.1.1 Choice of the kind of saliency map

This step corresponds to the choice of the more relevant kind of saliency map for each shot of the video to summarize. Indeed, static saliency maps highlight the textured and contrasted objects and are then adapted to describe the shots where motions are weak. At the opposite, dynamic maps highlight moving objects and are adapted for important motions of camera or objects. An attention curve (as defined by Ma et al.) is computed for each kind of saliency map. The attention curve with the higher standard deviation on amplitude corresponds to the kind of map able to give more information

and so this is this kind of map that is chosen. Experiments show that this criteria of choice gives the best final results.

##### 3.1.2 Keyframes selection

For the second step, the purpose is to build a summary with a reduce number of frame, but sufficiently different to represent the whole video. For that we propose to generate an “attention variation curve” on which keyframes are selected. To highlight changes during a video, a difference  $D_k$  of saliency map is computed:  $D_k = |M_k - M_{k-i}|$  where  $M_k$  is the saliency map of the frame  $k$  and  $i$  the parameter which defines the difference between current map and the  $i^{\text{th}}$  previous one. Experiments showed that  $i=10$  allows to have frames different enough for seeing changes and to take into account a rapid variation of saliency maps (to lighten when there are changes and to darken when there are no changes). The attention variation curve is then obtained by computing the average of grey level on each difference map  $D_k$ .

The selection of keyframes is done by detecting on the attention variation curve (example figure 2), the significant increase with regard to previous neighbours. An adaptive threshold is used for that purpose. The threshold is given by the expression  $\mu + 3\sigma$  where  $\mu$  is the average on a sliding window and  $\sigma$  is the standard deviation on the same window. This threshold takes the past into account, and permits to judge if the current frame is different enough from the previous ones. Tests have been made on shots with window's length of 10, 25, 50, 75 and 100 frames. Best results are obtained with a window of 50 frames (i.e. 2 seconds of video). Thus the threshold considers not too much or too few information about the past. A frame is selected each time the variation attention curve crosses the threshold.

##### 3.1.3 Elimination of redundant frames

A summary has been obtained for each shot. The summary was made trying not to forget events in the video and not taking neighbours frames, but frames with similar content can still be selected. A post-processing to eliminate redundant frames is proposed (figure 4). This process considers the static saliency map of each keyframe. A comparison between the saliency maps of the current frame and the previously selected one is done by computing their absolute difference. Then the average of the grey level of this absolute difference is calculated. If this value is below a threshold, the frames are too similar and so only the frames with the highest value on the attention curve are kept. Tests permit to choose the threshold equal to 0.11 and to be sure of its robustness.

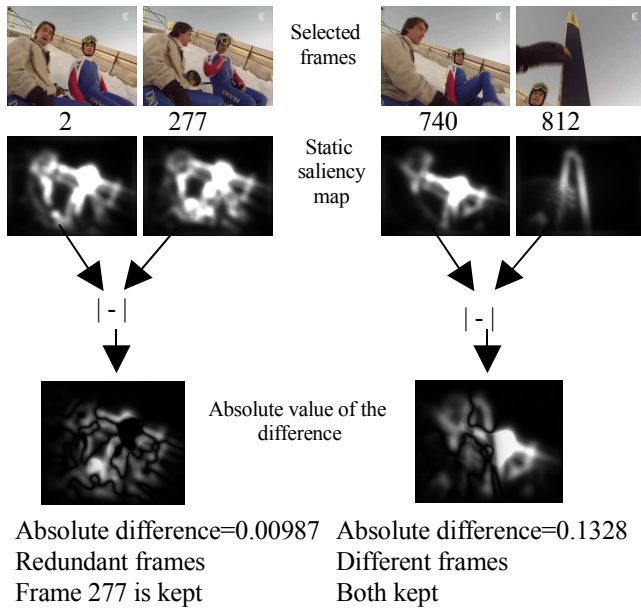


Figure 2 – Application of the redundant frame elimination process to two relevant cases.

An example of summary produced by this method is given figure 3. This method is used on long shot but an adaptation has been made for short shots.

### 3.2 Particular case of short shots

The particular case of short shots is considered here. A shot is an unstopped portion of video with or without continuous camera motion. We define a short shot as a shot with 100 frames or less, or 4 seconds of video at a video rate of 25 frames per second. All the frames of a short shot are close temporally. With the continuous variation of frames in video, all the frames of a 4 seconds shot have close content. Only one keyframe is sufficient to summarize that kind of shot. This keyframe is selected by taking the maximum of the attention curve using static saliency map. Static saliency maps are used because in 4 seconds or less the content of the shot do not change a lot and so the static saliency maps which highlight texture and contrast are more efficient than the dynamic ones which highlight motion.

## 4. SUMMARY EVALUATION METHOD

Different evaluation method exists. A possibility is to ask subject to choose between two summaries [7]. In this paper

the automatic summary is evaluated by comparing it to a “ground-truth” summary called reference summary.

### 4.1 Reference summary

To evaluate the summary previously generated, we compare it to a reference summary. To obtain this summary subjects are asked to watch the shots of a video and to make their ideal summary. They are told to make a summary, with from 1 to 3 keyframes for each shot. For a shot the number  $N$  of keyframes for the reference summary is given by the median of the number of frames selected by all the subjects. According to the number of frame chosen by a subject for a given shot a weight is given to each one of these frames. The more frames a subject selects the less their weight are. For each frame, the weight are summed for all the subjects. The reference summary keyframes are selected by taking the  $N$  frames with the highest weight.

### 4.2 Comparison method

The comparison between the automatic summary, called candidate summary, and the reference summary is carried out in 4 steps (figure 5).

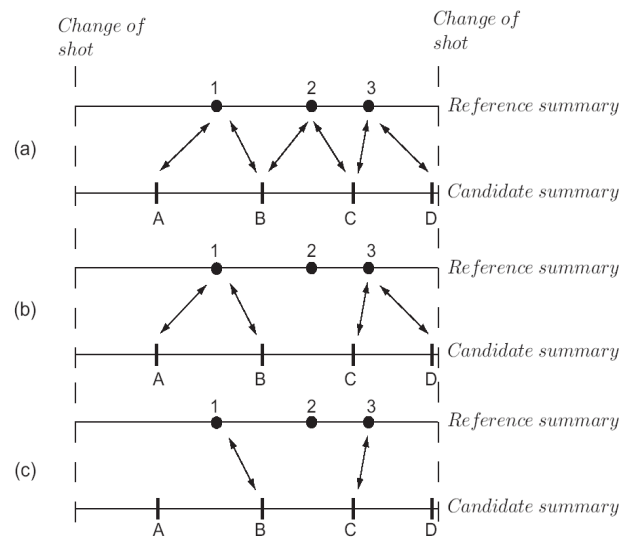


Figure 5 – three first step of the comparison method

The first step consists in associating frames of the candidate summary to frames of the reference summary. Each frame of the candidate summary is associated to two frames (at maximum) of the reference summary, the previous one and the

next one. During the second step frames of the candidate summary are only associated to the temporally closest frame of the reference summary. The third step selects only the frames of the candidate summary that are the closest with the one of the reference summary. The fourth step compares frames of the candidate summary to their associated frames on the reference summary by a colour histogram difference. The frames are temporally continuous; it is thus rather improbable to have two similar histograms with different contents.

## 5. EXPERIMENTAL RESULTS

The summarization method has been tested on 3 videos: an educative program, TV news and a series. All these videos represent 90 shots from 19 to 1468 frames.

The summary evaluation is done using the comparison method describe in section 4 and the criteria of recall R, precision P and their harmonic average F1:

$$R = \frac{N_f}{N_r} \quad P = \frac{N_f}{N_c} \quad F1 = 2 \frac{R \times P}{R + P}$$

where  $N_f$  is the number of frames of the candidate summary which correspond with frames of the reference summary,  $N_r$  the number of frames of the reference summary, and  $N_c$  the number of frames of the candidate summary.

Two other methods of summarization (a random summary and a method which selects a keyframe in the middle of each shot) are also compared to the reference summary. Results are given in table 1. For example if we consider the educative program and the summary using attention model (n°3). The recall (respectively the precision) shows that among the 24 (resp. 29) frames of the reference summary the automatic summary retrieves 15 frames.

The summary with visual attention model gives the best results. But the difference with the summary taking a keyframe in the middle of each shot is not high. This is explained by the fact that most of the shots are short shots (less than 100 frames). As we have already seen, the frames of short shots are temporally close and because of the slowness of the variations in video they are semantically close. Thus the two methods give similar results for short shots. The proposed method is more efficient on long shots.

## 6. CONCLUSION

In this paper we have presented a summarization method using a visual attention model. This method uses saliency map to compute changes and selects keyframes when they are changes in the video. This summarization method has been tested on three video of different length and content. For that purpose a reference summary and a comparison method were made-up. Results are quite good and improve with the shot length's.

The proposed summarization method has been carried out shot by shot. It could be extended to global summary of the video for reducing the redundancy and being more compact.

## REFERENCES

- [1] Y. Li, T. Zhang, D. Tretter, "An overview of video abstraction techniques", *HPL-2001-191*, 2001.
- [2] H. J. Zhang, J. Wu, D. Zhong, S. W. Smoliar. *An integrated system for content-based video retrieval and browsing I. Pattern recognition*, vol. 30, No.4, pp. 643-658, 1997
- [3] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Transaction on pattern analysis and machine intelligence*, vol.20, pp. 1254-1259, Nov. 1998.
- [4] L. Itti, N. Dhavale, F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention", in *SPIE 48<sup>th</sup> Annual International Symposium on Optical Science and Technology*, vol. 5200, pp. 64-78, 2003.
- [5] Y. Ma, X. Hua, H. Zhang, "A generic framework of user attention model and its application in video summarization", *IEEE Transaction on multimedia*, vol. 7, PP. 907-919, 2005.
- [6] M. Guironnet, N. Guyader, D. Pellerin, P. Ladret, "Static and dynamic feature-based visual attention model: comparison with human judgement", in *Pro EUSIPCO 2005, Antalya, Turkey*, Sept. 2005.
- [7] S. Corchs, G. Ciocca, R. Schettini, "Video summarization using a neurodynamical model of visual attention", in *IEEE 6<sup>th</sup> Workshop on Multimedia Signal Processing*, Sienna, Italy, pp. 71-74, Oct. 2004.

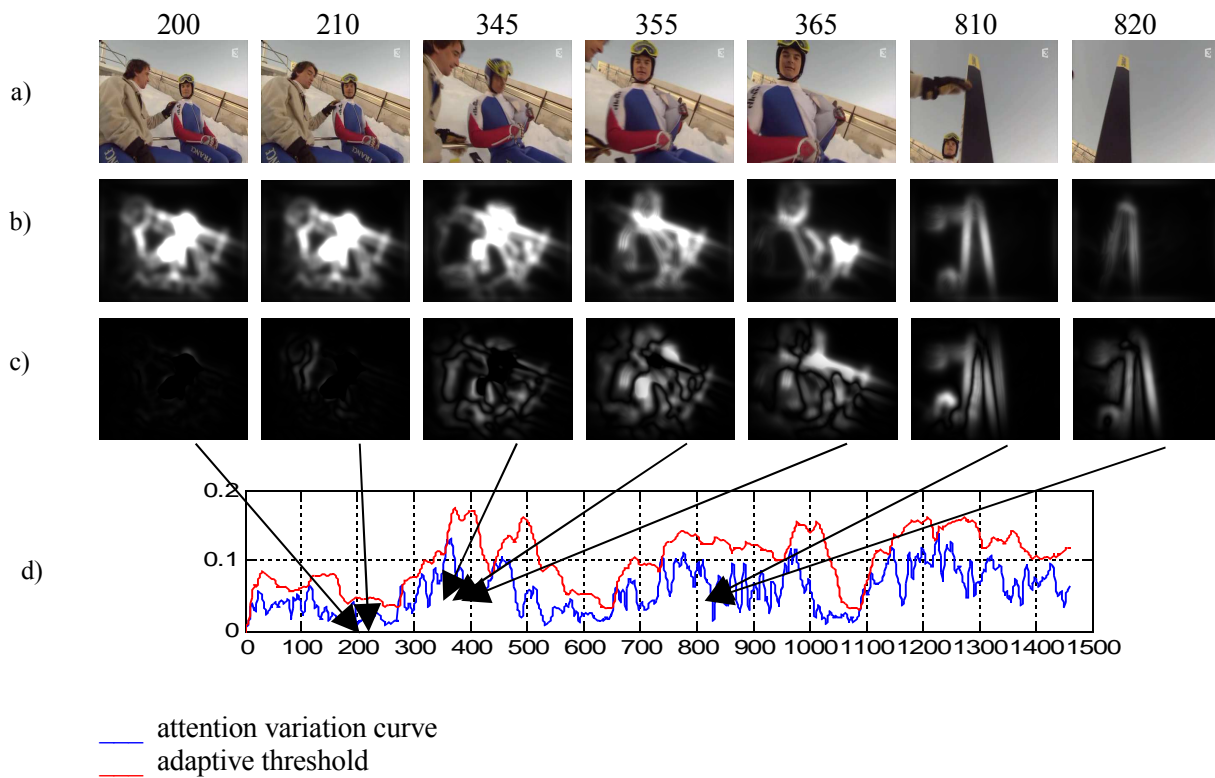


Figure 4 – Summary of the shot described in figure 3a.

summary	educative program			TV news			series		
	R	P	F1	R	P	F1	R	P	F1
n°1	62 (15/24)	40 (15/37)	49.1	83 (46/55)	50 (46/91)	63.0	80 (24/30)	40 (24/59)	53.9
n°2	50 (12/24)	60 (12/20)	54.5	63 (35/55)	83 (35/42)	72.1	73 (22/30)	78 (22/28)	75.8
n°3	62 (15/24)	51 (15/29)	56.6	78 (43/55)	70 (43/61)	74.1	80 (24/30)	75 (24/32)	77.4

Table 1: Results for 3 summarization methods on 3 videos (methods: n°1 random summary, n°2 summary selecting one frame at the middle of each shot, n°3 summary with the visual attention model ).