Spatio-temporal Attention Model for Video Content Analysis

Mickael Guironnet, Nathalie Guyader, Denis Pellerin, Patricia Ladret

Spatio-temporal Attention Model for Video Content Analysis

Mickael Guironnet*, Nathalie Guyader†, Denis Pellerin* and Patricia Ladret*
*Laboratoire des Images et des Signaux
46 avenue Felix Viallet, 38031 Grenoble, France
Email: {guironnet,pellerin,ladret}@lis.inpg.fr
†Department of Psychology, University College London
Gower Street London, England
Email: nguyader@yahoo.fr

***Abstract*— This paper presents a new model of human attention that allows salient areas to be extracted from video frames. As automatic understanding of video semantic content is still far from being achieved, attention model tends to mimic the focus of the human visual system. Most existing approaches extract the saliency of images in order to be used in multiple applications but they are not compared to human perception.**

The model described here is achieved by the fusion of a static model inspired by the human system and a model of moving object detection. The static model is divided into two steps: a "retinal" filtering followed by a "cortical" decomposition. The moving object detection is carried out by a compensation of camera motion. Then we compare the attention model output for different videos with human judgment. A psychophysical experiment is proposed to compare the model with visual human perception and to validate it. The experimental results indicate that the model achieves about 88% of precision. This shows the usefulness of the scheme and its potential in future applications.

I. Introduction

The quantity of audiovisual information has increased with the arrival of Internet and digital television. The need to analyze video semantic content has appeared to serve in many applications: video summary, video browsing, and video retrieval... Automatic understanding of semantic content is still far from being achieved in spite of the significant advances of computer vision and image processing.

Attention models have been introduced to solve the limits of current algorithms. Without understanding the full video content, attention models tend to capture the focus of the visual human system. Attention is a neurobiological concept that represents the capacity of humans to concentrate and focus on things such as a single object, a landscape... Thus, the understanding of attention processes should facilitate scene analysis and help video content analysis by reducing the number of objects or area to be analysed.

The first attention model was proposed by Koch and Ullman [1] in 1985. Then, Itti et al. [2] defined a visual attention map, which is the combination of different maps dedicated to different low level features (orientation, intensity and colour). Other authors, inspired by the human visual system, created more elaborated models. Chauvin et al. [3] proposed a model inspired by the retina and the primary visual cortex cell functionalities. Then, attention was introduced into videos to try to fill the semantic gap. The methods also tried to exploit the temporal component. Ma et al. [4] defined a user attention model based on a motion vector field extracted from MPEG stream. This approach was used for video skimming. New systems appeared by combining maps of static and dynamic visual attention. Bollmann et al. [5] introduced the detection of moving objects to the static features (symmetry, orientation and color analysis). Ho et al [6] presented an attention model based on three levels of features: low level (intensity and color), medium level (motion) and high level (face detection). The attention model defined in [7] aimed at simulating eye saccades. Two applications were considered: virtual humanoid perception and automatic video surveillance. In [8], the authors proposed an attention model based on many visual features (color, orientation and intensity) but also on face and text detection for adapting image size. In general, these models are not compared to human perception. They are directly used through various applications like video summarization, encoding, watermarking or surveillance.

This article presents a new visual attention model. It relies on the fusion of a static model inspired by the human system and a model of moving object detection in a scene. The static model is based on retinal filtering followed by a bank of Gabor filters. The moving object detection is carried out by compensation for camera motion. Once the visual spatio-temporal attention model was built, a psychophysical experiment allowed us to validate the proposed model. The main contributions of our work are a new user attention model and the building of an experiment to judge the effectiveness of the method.

II. Attention model

In this section, we describe our attention model. This model extracts the salient areas from videos. It is divided into two parts: a static and a dynamic one.

A. Static part of the model

This part is inspired by biology and functionalities of human visual system cells (from the retina to the primary visual cortex). This part of the model concerns each frame of

the videos.

Retinal filtering

At the first level of information processing, the retinal photoreceptors carry out an adaptive compression process followed by high-pass filtering [9]. This results in contrast equalization of the image, providing a relative insensitivity to local illumination variations. This pre-processing is interesting for extracting saliency because it is invariable to some modifications, such as luminance variations between images. Then, the parvocellular pathway provides a spatial high-pass filter (known to whiten an image's frequency spectrum) that compensates for the 1/f image amplitude spectrum.

Primary visual cortex

Primary visual cortex cells are sensitive to visual signal orientations and spatial frequencies. Here, we chose to model simple cell receptive fields. These cells are sensitive to stimuli having a certain orientation and a certain frequency with a specific position in the visual field, which is modelled by a two-dimensional Gabor filter. A Gabor function is defined by a gaussian with spatial extents σ_x and σ_y modulated by a complex exponential with frequency f in a direction θ. We carried out this filtering by directly multiplying the retina output image with the Gabor filter in the Fourier domain. Before achieving the Fourier transform, we multiplied the image by a Hanning window to remove edge effects. We chose here to decompose each image of the video using thirty-two Gabor filters (four different spatial frequencies and eight different orientations). So, we obtained thirty-two maps (thirty-two images), depending on the frequency and the orientation of the original image, for each frame of the video.

Interactions

A neuron is, by definition, a contact cell. Thus, the response of a cell is always dependant on a neuronal environment. Then, the neuron's activity is modeled by the visual field neighborhood and so is dependent on lateral connections. With regard to the orientations, the computed interactions preferentially connect neurons devoted to the same orientation. These interactions, which symbolize both excitatory and inhibitory connections, are modeled by a linear combination of the simple cells (fig 1):

$$E_{int}(f_i,\theta_j) = \sum_{k,l} w_{k,l} \cdot E(f_{i+k},\theta_{j+l}), w = \begin{pmatrix} 0 & -0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & -0.5 & 0 \end{pmatrix} \quad (1)$$

Figure 1 illustrates the interactions. In this example, the Gabor filter with weight 1 in direction $\pi/4$ interacts with its neighbours. The Gabor filters with the same direction symbolize excitatory and those with different directions represent inhibitory.

The output of this stage is composed of 32 maps for each image of a video. These maps put into relief the image energy in function of the spatial frequency and the orientation of the signal in the original image and take into account the interaction between the orientation maps.

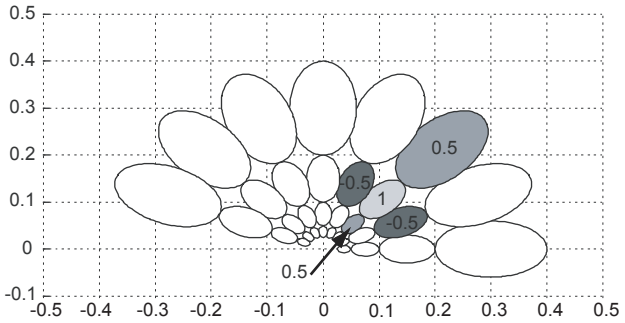

Fig. 1. Example of interactions. The Gabor filter with weight 1 in direction $\pi/4$ interacts with its neighbours.

Static saliency map

We extract a static saliency map for each image as the sum of the 32 energy maps described above:

$$E_f = \left| \sum_{i,j} E_{int}(f_i,\theta_j) \right| \quad (2)$$

The regions having the highest energy are considered to be salient. Figure 2 shows examples of static saliency maps. The content of the images is rather varied: a rugby match, a car chase and a bicycle race. We can observe on the bottom row that the energy is located on objects which seem to be salient.

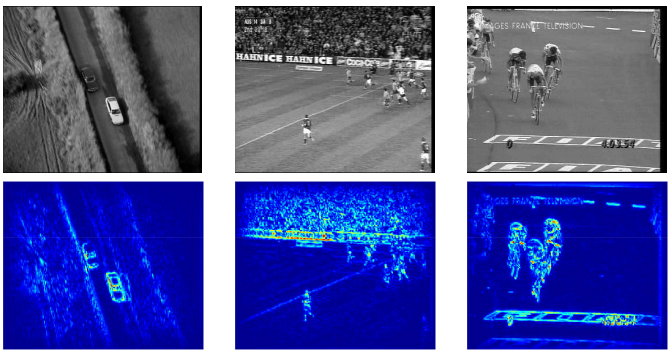

Fig. 2. Examples of static attention maps. Top row: video frame. Bottom row: the associated static attention map.

B. Dynamic part of the model

The dynamic part of the model detects the moving objects in a scene. In fact, we assume that the location where something moves is salient. Next, it is necessary to estimate camera motion. We use the 2D motion estimation algorithm developed in [10]. This algorithm provides the dominant motion between two successive frames. A 2D parametric motion model between two successive frames is then defined and a robust multiresolution estimation of parametric motion models is

carried out. We chose affine motion model to represent the camera motion.

$$\begin{cases} v_x = a_1 + a_2 \cdot x + a_3 \cdot y \\ v_y = a_4 + a_5 \cdot x + a_6 \cdot y \end{cases} \quad (3)$$

Once we had the coefficients $[a_1, ..., a_6]$, we computed the motion compensated frame. Compensation of camera motion was formed by bilinear interpolation $I_c(x, y, t+1) = I(x + v_x, y+v_y, t+1)$. The previous frame was then subtracted from the motion compensated frame to generate Displaced Frame Difference (DFD):

$$DFD(x, y, t) = I_c(x, y, t+1) - I(x, y, t) \quad (4)$$

Finally, the absolute value of DFD informs about regions that do not follow camera model and corresponds to displacement of objects. Figure 3 shows examples of object detection. We can see that the rugby players are correctly detected as well as the cars and the cyclists.

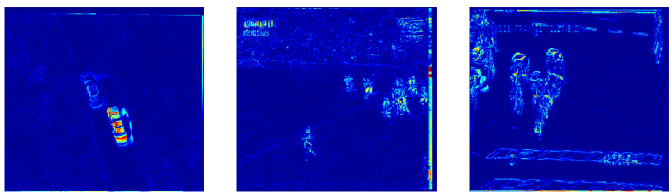

Fig. 3. Examples of moving object detection. Images represent the absolute value of Displaced Frame Difference.

C. Spatio-temporal attention model

Before dealing with map fusion, it is necessary to carry out a temporal filtering of each map. Indeed, the maps are computed locally, either on one frame (for the static model) or on two successive frames (for dynamic model), and the salient regions must be temporally coherent inside a window of duration L. The temporal continuity of the video prevents the appearance of salient areas on one or two images only (two frames correspond to $2/25 = 0.08s$). This is why a median filtering of width L is carried out. In our experiment, the size of window L equals five frames. Since the maps do not have the same magnitude, a standardization stage is necessary before carrying out the map fusion. This stage is carried out by the following:

$$S_n = \begin{cases} S/T_h & \text{if } S < T_h \\ 1 & \text{if } S \geq T_h \end{cases} \quad (5)$$

where S is a static or dynamic saliency map and T_h is a predefined threshold (25 in the two cases).

Once the maps have been normalized, a fusion stage is achieved to combine all maps into a final saliency map. The fusion is performed using the max operator which can be interpreted as an "or" logical operator. Thus, the final map contains static and dynamic information. Finally, the map obtained is a gray image with a higher value for salient zones.

Thanks to image processing techniques, we detect the regions of attention. The following steps are achieved: thresholding, morphological operation (close and open), region selection. In the last step, we determine the regions according to 4-connected neighbourhood. The regions with area lower than a threshold are removed. Finally, the remaining regions are selected and defined as masks. If the number of masks is greater than five, we keep only the five biggest masks. Figure 4 illustrates the fusion of maps and the selection of attention masks: the cars, crowd and players, and the cyclists.

Fig. 4. Examples of spatio-temporal attention masks.

III. Experimental results

In order to test and to validate our model, we carried out a psychophysical experiment. The goal of the experiment was to know if the areas defined as salient by the model are indeed salient. We tried to compare the model with human perception.

A. Method

Subjects

Sixteen naive subjects underwent the experiment. All subjects had normal or corrected to normal vision.

Stimuli

The subjects saw nine different videos displayed in the middle of the screen with a frequency of 25 images per second. Videos were composed of 288x352 pixel-images in 256 gray-levels. For one randomly selected image of one video, we associate to it the spatio-temporal attention map using the presented model. In order to test if the salient areas provided by the model are in agreement with the human visual perception we take exactly the same image and the same masks but we apply the masks to random positions as shown in figure 5. The principle is as follows: the mask of model having highest area is first randomly moved in the image; the second mask having highest area is then moved but without possible overlapping with the other mask and so on... Finally the two images (fig 5) have the same masks but at different places in the image.

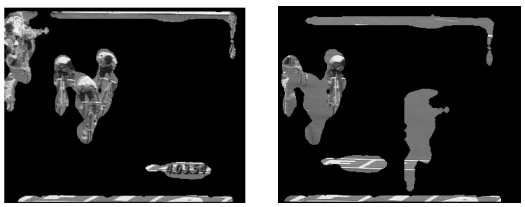

Fig. 5. Example of the target images. The left image is the output of the presented model. The right one is the same image with the same masks but placed in random positions.

Procedure

The experiment was processed with a computer with a Pentium III processor. The stimuli were presented on a 21" screen

(Mitsubishi Diamond Pro2020u) with a resolution of 1024 by 768 pixels and a frame rate of 100 Hz. Subjects were placed at a distance of approximately 50 cm from the screen. Figure 6 describes the events for one trial: a fixation point appeared (here a small black cross) in the middle of the screen for two seconds, followed by a 1.2s video still in the middle of the screen. Then, two images were presented symmetrically in the middle of the screen. These images belonged to the previous video and were masked in different ways: one following the model and the other with random position masks. The subject had to choose which one seemed to him to be the closest to the video. The selected image should have represented the best video content. His response and the reaction time were measured with a response box and E-Prime software. His answer had to be given as quickly as possible.

Each video appeared four times, with two different target images in the two possible positions on the screen. So in two cases the same images are used and in one case the image provided by the model is on the right side of the screen and in the other case is on the left. This allows us to have more answers for one condition and to see if a subject gave his answer randomly. The experiment is divided into three phases. Each phase contains three videos and so twelve random trials. During one experiment, each subject answered 36 trials.

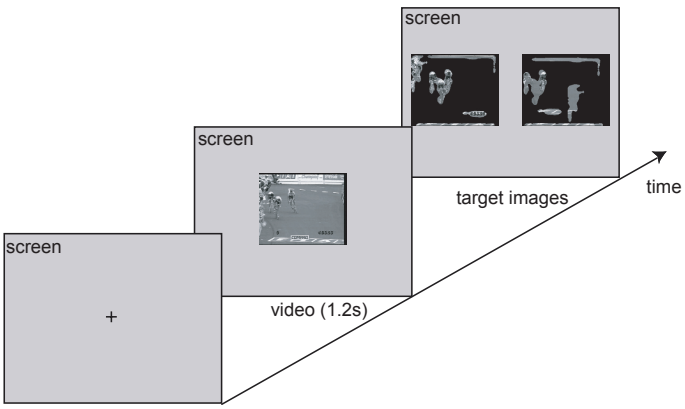

Fig. 6. Experimental design: one trial sequence is illustrated. First, a video appeared for 1.2s, and then two images appeared in the middle of the screen. The task was to choose the image that is closest to the video.

B. Results

For the analysis we only kept fourteen subjects (two subjects had random responses). We measured the percentage of correct responses per subject (the correct response is in the case where the subject chose the model mask). Over all subjects the mean correct response percentage is 88% with a mean standard deviation of 5%. As we expected, for all subjects the model masks correspond more to the video than the random masks.

We can refine these results. For some masks there is an overlapping between the model masks and the random ones. So we should find fewer correct answers when the model masks are overlapped with random masks rather than when the masks are separated. So we added a condition: when masks had more than 50% of overlapping and when masks had less than 50% of overlapping. We made an analysis of variance (ANOVA) for the percentage of correct response as a function of these two conditions. The overlapping influences the correct response ($F(1, 13) < 0.001$). So the percentage of correct responses is lower when the model masks and the random masks are overlapped, which consolidates the model.

IV. Conclusion

We have presented a spatio-temporal attention model. It relies on the fusion of a static model inspired by the human system with a model of moving object detection. A psychophysical experiment was proposed to judge the effectiveness of the model. The proposed model provides good results with a precision of 88%. These results are promising. In addition, the model can be used in many applications such as video indexing, summarization, watermarking and surveillance. One of the future works would be to use this model to provide a video summary and to use an experimental paradigm to test the efficiency of the method.

Acknowledgment

The authors would like to thank Christian Marendaz (Laboratoire de Psychologie et NeuroCognition, Grenoble, France) for having received us and let us use the material for the experiments. The authors also thank the Vista Research team at Irisa/Inria Rennes for the use of the Motion2D software.

References

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *In Human Neurobiology, Springer-Verlag*, pp. 219–227, 1985.

[2] L. Itti and C. Koch, "Target detection using saliency-based attention," in *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified), Utrecht, The Netherlands, RTO-MP-45 AC/323(SCI)TP/19*, Jun 1999, pp. 3.1–3.10.

[3] A. Chauvin, J. Herault, C. Marendaz, and C. Peyrin, "Natural scene perception: visual attractors and image processing," *Connectionist Models of Cognition and Perception, Proceedings of the Seventh Neural Computation and Psychology Workshop, World Scientific Press*, pp. 236 – 245, 2002.

[4] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the tenth ACM international conference on Multimedia*. ACM Press, dec 2002, pp. 533–542.

[5] M. Bollmann, R. Hoischen, and B. Mertsching, "Integration of static and dynamic scene features guiding visual attention," in *Mustererkennung 1997, 19. DAGM-Symposium*. Springer-Verlag, 1997, pp. 483–490.

[6] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu, "A user-attention based focus detection framework and its application," in *Proceedings of the fourth International Conference on Information, Communications and Signal Processing and Fourth Pacific-Rim Conference on Multimedia (ICICS-PCM'2003)*, vol. 3, dec 2003, pp. 1315 – 1319.

[7] N. Courty, E. Marchand, and B. Arnaldi, "A new application for saliency maps: Synthetic vision of autonomous actors," in *International Conference on Image Processing (ICIP'03)*, Barcelona, Spain, September 2003.

[8] L.-Q. Chen, X. Xie, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "Image adaptation based on attention model for small form factor devices," in *the 9th International Conference on Multimedia Modeling*, jan 2003, pp. 483–490.

[9] W. H. A. Beaudot, "Sensory coding in the vertebrate retina: towards an adaptive control of visual sensitivity," *Network: Computation in Neural Systems*, vol. 7, pp. 317–323, 1996.

[10] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models in complex image sequences," in *Proc. 7th European Conf. on Signal Processing, Eusipco'94*, Edinburgh, Scotland, September 1994.