IEEE INT. CONF. ON IMAGE PROCESSING, ICIP'03
BARCELONA, SPAIN, SEPTEMBER 2003

A NEW APPLICATION FOR SALIENCY MAPS: SYNTHETIC VISION OF AUTONOMOUS ACTORS

Nicolas Courty, Éric Marchand, Bruno Arnaldi

IRISA - INRIA Rennes
Campus de Beaulieu,
35042 Rennes Cedex, France
Email nicolas.courty@irisa.fr

ABSTRACT

We present in this paper a new and original application for saliency maps, intending to simulate the visual perception of a synthetic actor. Within computer graphics field, simulating virtual humans has become a challenging task. Animating such an autonomous actor within a virtual environment requires most of time the modeling of a perception-decision-action cycle. To model a part of the perception process, we have designed a new model of saliency map, based on geometric and depth information, allowing our synthetic humanoid to perceive its environment in a biologically plausible way.

1. INTRODUCTION

Simulating virtual humanoids has become within computer graphics field a vast research area. The main goal is to be able to simulate human beings evolving into unknown environments. The applications of such models can be either video games, conversational agents, simulators. In order to simulate autonomous behaviors for these humanoids, a perception-decision-action cycle is needed. Perception can be achieved through a direct access to the world database or a subset of this data-base, or through synthetic vision [1, 9, 10, 7]. We will mainly focus on those approaches as far as they are concerned with image processing. There have been several previous work for modeling simulated visual perception as a treatment of an image rendered from the point of view of the synthetic actor. Blumberg used an image-based motion energy designed for obstacle avoidance dedicated to his virtual dog [1]. Terzopoulos and Rabie built an iterative pattern-matching scheme for object recognition based on the comparison between color histograms of the pre-rendered images of virtual objects, and the current images [10]. Noser [9] used a false-coloring rendering technique, as well as Kuffner [7] to determine a list of visible objects in order to represent visual memories of the different actors. None of these approaches have considered properties of the human visual system.

Chopra [2] distinguished different types of visual tasks. Some of them, relevant to endogenous stimulies, are part of a conscious process. In our system, those aspects have been treated aside. Our main interest is to be able to model spontaneous looking, i.e.to be able to determine, without any knowledge about the environment, where the attention of the humanoid will be drawn. This process is performed using saliency maps. Saliency maps have been widely used in the computer vision field to solve the attention selection problem [12, 4, 6]. The purpose of the saliency map is to represent the conspicuity (as a scalar value) of each locations of the visual field. According to such a map, it is possible to guide the look of the humanoid, depending on the spatial distributions of those locations. The main idea of this paper is to be able to build a saliency map from an image rendered from the point of view of the humanoid. Figure 1.a shows a representation of such a process.

The next sections of this paper intend to present the architecture of our model and the first results we obtained using such a technique into virtual environments.

2. ARCHITECTURE OF THE MODEL

In order to test our idea we have designed a simple and original model of saliency map based on two different feature maps. Those maps give information about spatial frequencies and about the depth of objects. It is possible to enhance this system with other feature maps (such as color or intensity maps). We choose to use a depth map as far as depth plays an important role in the attention selection process [11], and as this type of information is rarely used in classical 2D image processing. Figure 1.b displays the combination of those feature maps and the construction of the saliency map. These different steps are described below.

Acquisition of the feature maps. The rendered image is directly given by the 3D engine of the application, as well

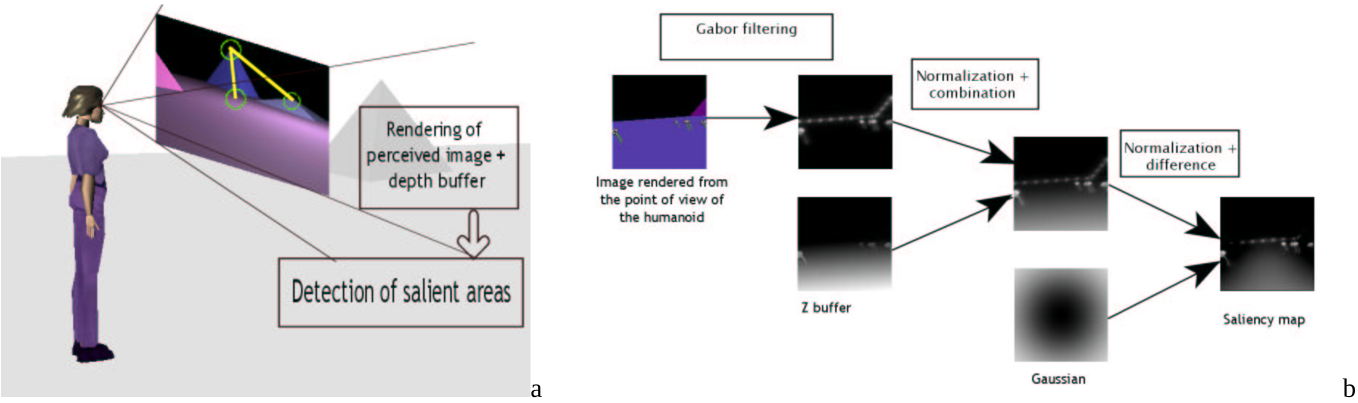
Rendering of perceived image + depth buffer
Detection of salient areas
a
Gabor filtering
Normalization + combination
Image rendered from the point of view of the humanoid
Normalization + difference
Z buffer
Gaussian
Saliency map
b

Fig. 1. (a) Using saliency map to perceive the environment. Green circles are conspicuous zones, and the yellow lines are a pathway between those zones (b) obtaining a saliency map based on spatial frequencies and depth information

as the depth buffer, also called Z-buffer (which is an essential information in the rendering process). The image corresponds to a given viewing pyramid parameterized to respect size and aperture of human vision (we designed this pyramid to respect an aperture of 120 degrees of human vision). The quality of the final image depends on the type of rendering algorithms. We will not discuss in this paper the influence of those algorithms in the final results, though we have considered this issue in our study. Let I be the rendered image from the view point of the eyes of the humanoid. In our implementation, we set the size of this image at 256×256 pixels, which has appeared to be sufficient regarding to the types of environment we have used. The corresponding depth buffer I_{depth} has the same size, and is composed of scalar values restricted to $[0 \cdots 1]$ (1 should be just in front of the humanoid), obtained from the perspective projection of the objects in the viewing pyramid. Those values are scaled to $[0 \cdots 256]$ using a simple energy normalization operator. Our hypothesis is that close objects should be more salient for the humanoid. Conversely, it is also possible to consider that objects at a particular range of depth are more salient regarding to focus considerations. In this case, the depth map should be filtered with a band-pass filter parameterized to enhance this particular range of depth.

2D Gabor filtering. Gabor filtering allows to get information about local orientation in the image. Bidimensional Gabor filters are part of a family of bidimensional Gaussian functions modulated by a complex exponential:

$$G_{f,\theta,\sigma}(x,y) = \frac{1}{2\pi\sigma^2} e^{(-\frac{x^2}{2\sigma^2} - \frac{y^2}{2\sigma^2})} e^{(j2\pi f(x\cos(\theta)+y\sin(\theta)))} \tag{1}$$

where f is the frequency, θ the orientation and σ the scale. These filters have the particularity to approximate the receptive field sensitivity of orientation-sensitive cells of the retina [8, 5]. We then convolute the rendered image with a bank of Gabor filters obtained from particular orientations ($\theta_i = \frac{i\pi}{n}$) and particular scales σ_i. If $I_{orientation}$ is the resulting orientation map, $I_{orientation}$ is given by:

$$I_{orientation} = \sum_{i=1}^{S} \sum_{j=1}^{O} I \star G_{f,\theta_j,\sigma_i} \tag{2}$$

with S the number of scales, O the number of orientations and $\star$ the convolution operator. The orientation map is then normalized with a traditional normalization operator. Let us note that it exists some fast computation algorithms for Gabor filtering. Applying such filters on synthetic images may cause irrelevant details to appear due to an insufficient quality of the rendering process. Typically, straight lines may appear crenellated. Using anti-aliasing can solve such types of problem (equivalent to a Gaussian filtering of the image). This shows the particular importance of the quality of the rendering process in the validity of the saliency map.

Simulating center-surround. Once the depth map and the orientation map have been computed, the two maps are normalized and combined into one map I_t:

$$I_t = N\left(N\left(I_{orientation}\right) + N\left(I_{depth}\right)\right) \tag{3}$$

if $N()$ is the energy normalization operator. We subtract to this map the image of a Gaussian I_{gauss} to simulate the center surround process, i.e.:

$$I_{final} = I_t - I_{gauss} \tag{4}$$

This Gaussian can be parameterized accordingly to the physical characteristic of the synthetic actor. We then obtain the final saliency map I_{final}.

3. HANDLING OUTPUTS OF THE TREATMENTS

Thanks to the saliency map, it is then possible to build a list of fixations. Those salient zones are defined as a local maximum of energy in the image. The process used to build

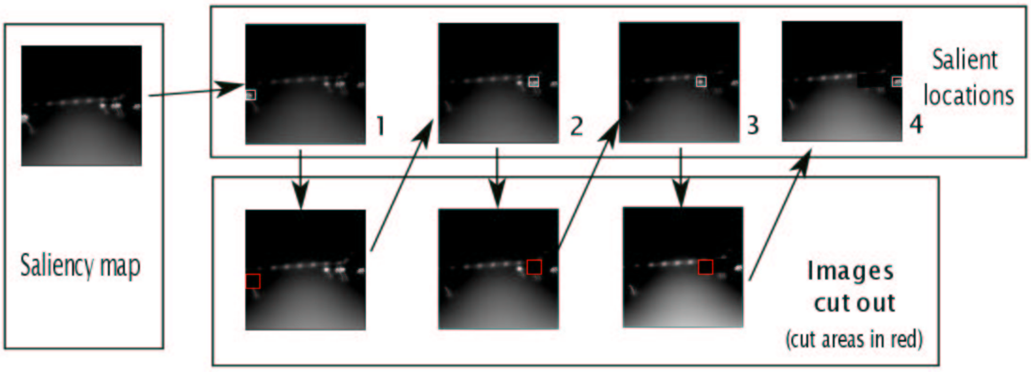

Fig. 2. Detecting a list of fixations; white squares show maximums of energy, red squares show cut out zones in the saliency map

such a list is depicted in Figure 2. First of all the global maximum of the map is determined through a simple sweep of the saliency map. Then the saliency map is cut out of a square zone whose size depends on the size of the viewing pyramid, to prevent one zone to attract all the fixations (inhibition mechanism). This process is then repeated until the saliency goes under a given threshold.

Connecting with the animation system. The list of fixations is composed of 2D points expressed in the image frame. We can now convert the locations of these pixels back into 3D world by inverting and applying the graphic pipeline rendering transforms. The 3D points are then given as inputs to our image-based animation engine [3], and an animation is generated.

4. RESULTS

We have tested our system on a virtual character wandering along the street into a virtual city. This animation was generated in real-time on SGI330 Linux PC, under our animation and simulation framework. Figure 4 show different salient maps computed along his path. Considering the Gabor filtering part, 8 orientations and 4 scales were used. A special thread was designed to compute those maps, and the time elapsed during its execution was around 100 ms, which is less than the time needed to gaze at the different conspicuous zones. One can observe that most of the important objects are detected : traffic signs, houses, sidewalks. Hence, at the end of the animation, our humanoid had looked at many different locations. In comparison, a traditional animation system would have required to set in the environment some predefined targets, which presupposes pieces of knowledge about the environment. The model of saliency map we designed is rather simple, and could be enhanced with other types of feature maps. Meanwhile, the resulting animation is quite interesting in the sense that it gives to our virtual character a lively, human-like behavior (see Figure 3 for snapshots of the animation).

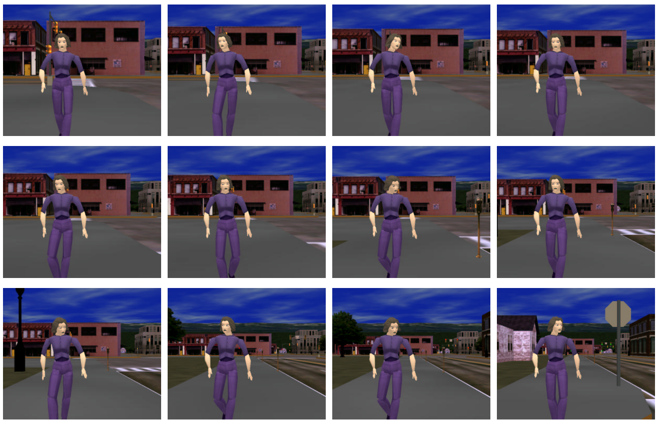

Fig. 3. Snapshots of the animation where our virtual characters walks along the sideway

Discussion. When the virtual character is standing still, the different salient maps are quite similar, resulting in the fact that from one map to the other, the humanoid is looking at the same locations. This shows that such a system needs to be coupled with a type of visual memory to avoid such configurations to happen. Moreover, validating our system is quite a difficult task, as far as a parallel with real experimentations can't be drawn. We intend to experiment it with photo-realistic rendered environment corresponding to real ones, and compare the results with human subjects.

5. CONCLUSION

In this paper we have presented a new and original application for saliency maps. We have designed a simple and original model of saliency map that is adapted to our com-

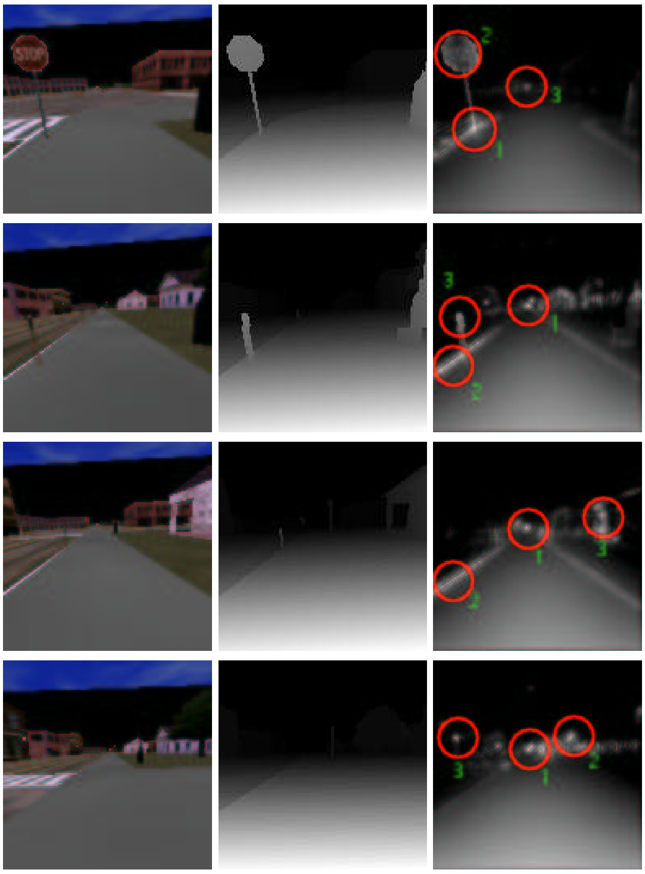

Fig. 4. Different saliency maps obtained during the wandering. First column shows the corresponding subjective view, second column shows the depth buffer and the last column is the saliency map with the location of three fixations

puter graphics problematic. While other approaches didn't take into account the properties of the human visual perception system, we have proposed a way to simulate in a more accurate way humanoids evolving into virtual environments. Information used in the image analysis process are of a new type (depth information for instance), and we think that such an application can be of use to improve attention selection models. One of the most important problems lies in the validation of our model regarding to simulation. Processing computer rendered images has revealed to depend on the quality of the rendering step. In our next work, we intend to apply this technique on photo-realistic rendered images of existing environments, and compare our results with real ones. Moreover, improvements can also be performed in the definition of the saliency map, notably thanks to the addition of other types of feature maps. Considering the temporal aspects, it seems to be of need to add a top-down attention selection process (based on memory) that can come up with the redundancy problem existing between two or more consecutive maps.

6. REFERENCES

[1] B. Blumberg and T. Galyean. Multi-level direction of autonomous creatures for real-time virtual environments. In *Proc. of SIGGRAPH 95, in Computer Graphics Proceedings*, pages 47–54, Los Angeles, Californie, August 1995.

[2] S. Chopra-Khullar and N. Badler. Where to look? automating attending behaviors of virtual human characters. In *Proc. of the 3rd Int. Conf. on Autonomous Agents (AA-99)*, pages 16–23, New York, May 1999.

[3] N. Courty, E. Marchand, and B. Arnaldi. Through-the-eyes control of a virtual humanoïd. In *IEEE Int. Conf. on Computer Animation 2001*, pages 234–244, Seoul, Korea, November 2001.

[4] S. Culhane and J.K. Tsotsos. An attentional prototype for early vision. In *Proceedings of Computer Vision (ECCV '92)*, volume 588 of *LNCS*, pages 551–562, Berlin, Germany, May 1992.

[5] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1169–1179, July 1985.

[6] L. Itti, J. Braun, D.. Lee, and C. Koch. A model of early visual processing. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[7] J. Kuffner and J.C Latombe. fast synthetic vision, memory, and learning models for virtual humans. In *Proc. of Computer Animation '99*, pages 118–127, Genève, Suisse, mai 1999.

[8] S. Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of Optical Society of America*, 70:1297–1300, 1980.

[9] H. Noser, O. Renault, D. Thalmann, and N. Magnenat-Thalmann. Navigation for digital actors based on synthetic vision, memory and learning. *Computer & Graphics*, 19(1):7–19, 1995.

[10] D. Terzopoulos, T. Rabie, and R. Grzeszczuk. Perception and learning in artificial animals. In *Proc. of the 5th Int. Workshop on Artificial Life : Synthesis and Simulation of Living Systems (ALIFE-96)*, pages 346–353, Cambridge, May 1997.

[11] J.M. Wolfe and G. Gancarz. Guided search 3.0. In *Basic and Clinical Applications of Vision Science*, pages 189–192, Dordrecht, Netherlands, 1996. Kluwer Academic.

[12] A. Yarbus. *Eye movements and Vision*. New-york : Plenum Press, 1967.