Modelling Visual Attention in VR: Measuring the Accuracy of Predicted Scanpaths

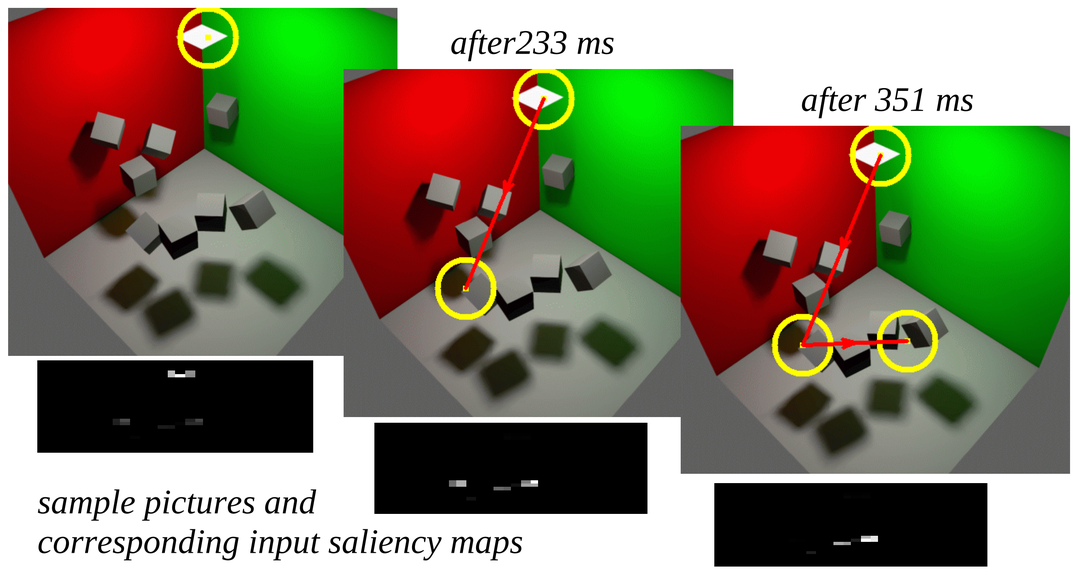
Gerd Marmitt and Andrew T. Duchowski

To increase the visual fidelity of Virtual Environments (VEs) recent strategies have emerged which combine view−independent rendering solutions with view−dependent perceptually−driven enhancements. Such techniques exploit the perceptual limitations of the Human Visual System and reduce the computational burden by e.g., focusing computational resources within highly salient regions (Regions of Interest, ROIs). The key to "just−in−time" view−dependent enhancements is the determination of instantaneous ROIs, which can be predicted by a computational model. In this paper a visual attention model developed by Itti et al. and previously used in VEs is compared to human viewing patterns in VR.

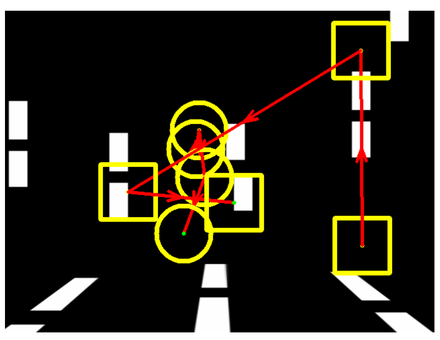
1. Visual Attention Modeling

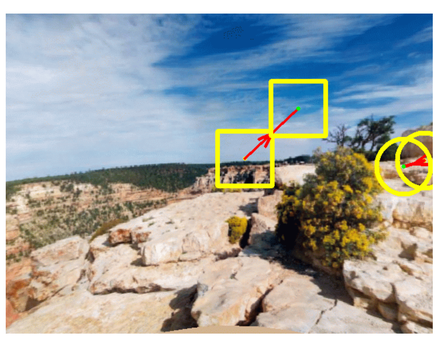
Itti et al.'s visual attention model is a bottom−up, task−independent algorithm which computes a saliency map from an input image based on color, orientation and intensity properties of the image. A winner−take−all neural network selects the most salient point within this map.

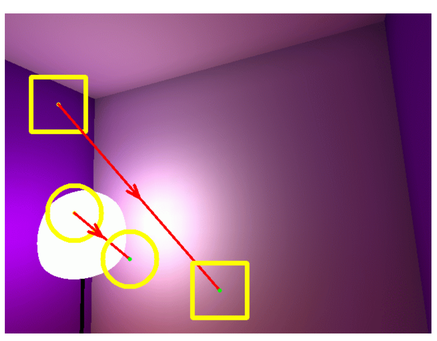
The example besides shows the first three steps of the model. The small b/w images are the saliency maps. White areas are predicted to be viewed by the human and will therefore be selected first. Note that the saliency map is computed once per image.



sample pictures and corresponding input saliency maps







scanpath comparison for all three types of envirnoments. Circles represent human scanpaths while rectangles represent the prediction of the attentional model for this particular image

S_s	Subj. 1		Subj. 2	
	Pict1	Pict2	Pict 1	Pict 2
S1P1	**R**	**I**	**L**	**G**
S1P2		**R**	**G**	**L**
S2P1			**R**	**I**
S2P2				**R**

structure of Y−Matrices (up) and Parsing−Diagrams(down) for commulating the results

	Same Subj.	Diff. Subj.
Same Image (SI)	**R**epetitive	**L**ocal
Diff. Image (DI)	**I**diosyncratic	**G**lobal
	S_s	Random

2. Comparision of Human and Artificial ROIs

Given two sequences of ROIs (scanpaths) two comparisons are possible: positional similarity and sequence similarity. For both it is necessary to convert the scanpaths into string sequences.

While path similarity does not consider the order of ROIs, sequence similarity uses the Levenstein distance to compare two strings. The final result, a parsing diagram, illustrates five different similarities. Most important are: idiosyncratic (same subject, different images), local (different subjects, same image) and global (different subjects and images).

3. Head−Based Analysis

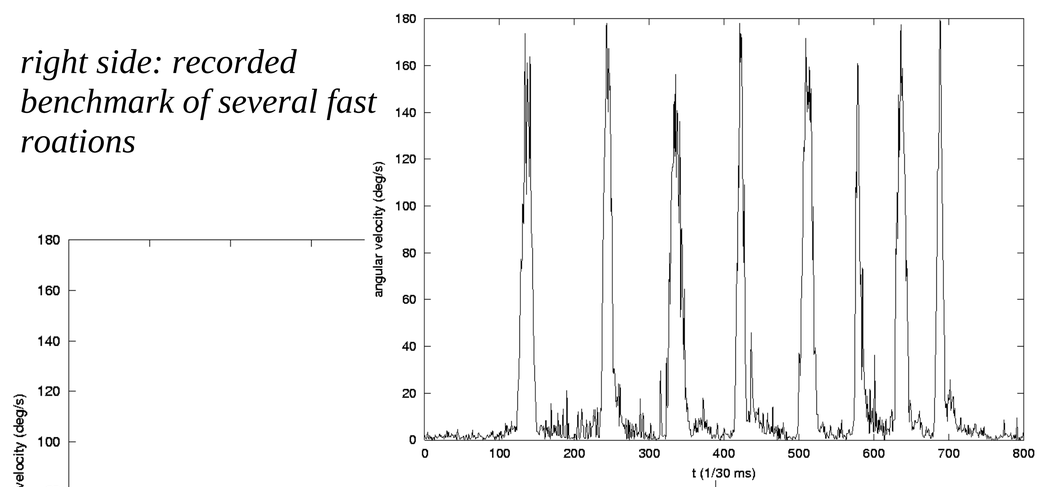
This analysis aims at isolating periods of immersion in the VE where the head (and hence image) is stable. The resulting sequence of image frames, averaged to a single image, is used to compare human and artificial scanpaths since both the attentional model and the comparison methods are based on still images.

To achieve this goal, Euler angles are recomputed from each captured data and periods of small changes (i.e., low velocity of the head) are grouped together. Each group is than compared separately.

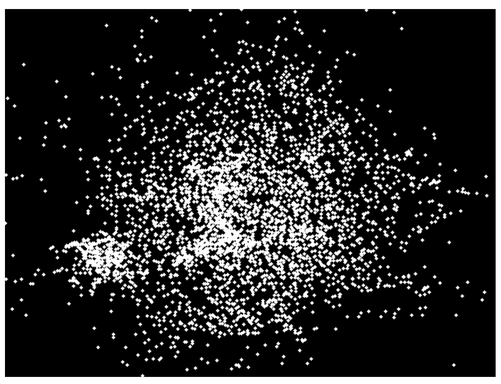
4. Time−Based Analysis

In this method, an image is taken every 10 ms from the captured data. If a human fixation is detected for this particular image, a special version of the attentional model also calculates one fixation. The distance between both fixations is than calculated in visual angles and averaged over the entire sequence.

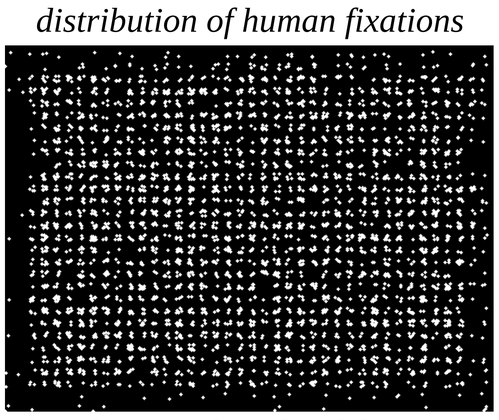
right side: recorded benchmark of several fast roations



left side: recorded benchmark of one slow head rotation



distribution of human fixations



distribution of artificial fixations

5. Results

To check the performance of the model, four cases were constructed, e.g., restricting the time or removing parts of the model. There was no significant difference between the 4 cases. The experimental design included a binocular eye tracker mounted in a typical HMD.

Nine subjects participated, each immersed in three different environments (of different complexity). In all three levels of VEs were presented, each with different content. The results of both head− and time−based analyses showed that the model performs poorly in Virtual Environments. In only 10 % of all comparisons the distance between human and artificial ROIs was acceptably small.

Improved prediction of human attention may nevertheless be possible if the model extended in two ways. First, the model should pay more attention the central region of the VR display (see human and artificial fixation distributions at left). Second, the model should be augmented with memory of the saliency map. This would decrease the computational cost since only new regions would be recomputed and an LRU strategy could be employed to improve prediction.

Level-0	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.007**
	0.162	0.104	0.006
		S_p	0.044

Same Image / Diff. Image	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.007**
	0.148	0.000	0.004
		S_s	0.000

Level-1	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.028**
	0.153	0.114	0.017
		S_p	0.031

Same Image / Diff. Image	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.028**
	0.151	0.004	0.016
		S_s	0.013

Level-2	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.007**
	0.197	0.081	0.015
		S_p	0.020

Same Image / Diff. Image	Same Subj. (human)	Same Subj. (model)	Diff. Subj.
	×	×	**0.007**
	0.191	0.004	0.013
		S_s	0.017

head (up) and time (down) based analysis for one case

Case 1

	L-0	L-1	L-2	L-0
100 %	48.30	42.70	43.38	46.69
90 %	43.10	36.75	37.31	41.01
80 %	36.99	32.70	32.45	36.44
70 %	34.23	29.24	28.81	33.51
60 %	30.97	25.75	25.42	30.42
50 %	28.08	22.61	21.57	26.09
40 %	23.76	19.39	18.24	22.01
30 %	18.23	16.09	14.65	18.57
20 %	14.31	11.38	10.18	13.35
10 %	4.01	2.86	2.03	3.56