In: Workshop on Organic Computing, P. Dadam, M. Reichert (eds.), Proceedings of the 34th GI-Jahrestagung, Vol. 2, 646–650, Ulm, September 2004.

Saliency Extraction for Gaze-Contingent Displays

Martin Böhme, Christopher Krause, Thomas Martinetz, and Erhardt Barth

{boehme, krause, martinetz, barth}@inb.uni-luebeck.de

Institute for Neuro- and Bioinformatics
University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

Abstract: Organic Computing is beginning to provide computer systems with organic and biological properties. We believe these systems will benefit from user interfaces that integrate the user more tightly with the system and optimize the way the user absorbs information. To this end, we propose *gaze-contingent interactive displays* that monitor and guide the user's gaze to make interaction with the system more effective and enjoyable.

A vital prerequisite for such a system is the ability to predict one or several salient locations that the user is likely to attend to in a dynamic display, with the goal of then modifying the display to influence the user's direction of gaze. In this paper, we present a structure-tensor-based saliency measure and a novel algorithm for extracting salient locations from saliency maps by using the mechanism of selective lateral inhibition. We assess the quality of the extracted locations by comparing them to the locations actually attended by test subjects.

1 Introduction

Organic Computing [Org] promises to simplify the process of configuring, maintaining and interacting with computer systems. Providing a computer system with organic and biological properties brings its behaviour closer to that of the human who interacts with it. We believe this is an excellent opportunity for applying new interaction technologies to fuse the user-computer system into an organic whole.

The display is an important part of most human-computer interfaces. So far, displays have been *passive* – they do not react to the way they are viewed, nor can they influence the way they are viewed. We propose the concept of a *gaze-contingent interactive display* that works in conjunction with an eye-tracker to adjust itself to the way it is being viewed and guide the user's gaze to optimize the way in which the user absorbs information.

The potential uses for such gaze-contingent displays are far-reaching. They include automobile applications, where the driver's gaze could be directed automatically towards an obstacle when sensors detect the danger of a collision; training applications in fields where a lot of visual information has to be absorbed and processed in a short amount of time (e.g. flight simulators, radiology); and reading-support systems that could increase the speed

and reduce the fatigue of reading as well as support persons with impaired vision [Ita].

If a system is to influence the user's direction of gaze, it must be able to predict one or several locations where the gaze will fall and then enhance or suppress the stimuli at those locations depending on the desired direction of gaze. In this paper, we will present a new method, based on the mechanism of selective lateral inhibition, for extracting salient locations from a saliency map. We apply this method to various saliency measures based on the structure tensor and assess the quality of the extracted locations by comparing them to the locations actually attended by test subjects in two test videos. Note that our approach takes dynamic aspects of the scene into account. While a lot of work has been done on computing saliency in static images (e.g. [IK01]), only a few specialized algorithms exist for dynamic saliency (e.g. [BMS02]). To our knowledge, no comprehensive study of dynamic saliency has been performed.

Section 2 describes the saliency measures and the salient location extraction algorithm. Section 3 presents the results obtained on two test video sequences. Section 4 summarizes our findings and discusses avenues for future research.

2 Method

2.1 Saliency Map Generation

Our approach to saliency is based on the concept of intrinsic dimensionality and is implemented using the structure tensor **J**, which is defined based on the image-intensity function $f(x, y, t)$:

$$\mathbf{J} = \mathrm{w} * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix}$$

where subscripts indicate partial derivatives and w is a spatial smoothing kernel that is applied to the products of first-order derivatives. The intrinsic dimension (iD) of f is n if n eigenvalues λ_i of **J** are non-zero. However, we derive the iD from the invariants of **J**, which are

$$\begin{aligned} H &= \mathrm{trace}(\mathbf{J}) && = \lambda_1 + \lambda_2 + \lambda_3 \\ S &= \mathrm{M}_{11} + \mathrm{M}_{22} + \mathrm{M}_{33} && = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K &= \det(\mathbf{J}) && = \lambda_1\lambda_2\lambda_3, \end{aligned}$$

where M_{ij} are the minors of **J** obtained by eliminating row $4-i$ and column $4-j$ of **J**. The iD is at least 1 if $H \neq 0$, at least 2 if $S \neq 0$, and 3 if $K \neq 0$. We will use these invariants as saliency measures because the information at a location (x, y, t) is less redundant if the iD is higher. The invariants of **J** are chosen since they are a straightforward method for estimating the iD. To extract salient locations on different spatial and temporal scales, we use a 4-level spatio-temporal Gaussian pyramid and compute the saliency measures on each level. Such a pyramid is constructed from the image sequence by successive blurring and subsampling.

2.2 Salient location extraction

In previous work [BDM03, BKBM], we extracted salient locations from the saliency map by applying a threshold of 0.5 of the maximum saliency in the map. For each connected region with saliency values above the threshold, we extracted one salient location by determining the location with maximum saliency. The robustness of this approach proved to be unsatisfactory. For example, if a small region in the saliency map has values substantially higher than the rest of the map, all of the map except for the high-saliency region will be suppressed.

For this reason, we developed a new, more robust approach, based on the mechanism of "selective lateral inhibition". The idea is that it does not make sense to have two salient locations generated closer together than a certain distance. Therefore, when a location has been extracted, we attenuate the saliency values of points around the location using a Gaussian kernel to inhibit the generation of further salient locations close to the existing location.

The following algorithm uses the mechanism of selective lateral inhibition to extract n salient locations from the image in the order of decreasing saliency:

$\mathcal{S}_1 = \mathcal{S}$
for $i = 1 \ldots n$ **do**
$(x_i, y_i) := \underset{(x,y)}{\operatorname{argmax}}\, \mathcal{S}_i(x, y)$

$$\mathcal{S}_{i+1}(x, y) := \begin{cases} \mathcal{S}_i(x, y) \cdot G(x - x_i, y - y_i) & x_i - W < x < x_i + W \text{ and} \\ & y_i - W < y < y_i + W \\ \mathcal{S}_i(x, y) & \text{otherwise} \end{cases}$$

end for

where $\mathcal{S}$ is the saliency measure (one of H, S or K), $G(x, y) = 1 - \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}}$ is an inverted radial Gaussian and W is the width of a window chosen such that $G \approx 1$ at the edges of the window. We thus repeatedly find the point with maximum saliency and then attenuate the saliency values around this point. The extracted locations are then just the (x_i, y_i).

3 Results

The quality of the salient locations was evaluated by comparing them to the locations attended by test subjects on two video sequences. The first sequence (30 seconds, 25 fps, 360×288 pixels) is synthetic, showing a square that moves from top left to bottom right. In addition, two other squares pop in and out at different moments. The second sequence (15 seconds, 25 fps, 352×288 pixels) shows traffic flowing across an intersection.

The monitor used had an image size of 40 by 30 cm at a viewing distance of 50 cm, spanning a horizontal field of view of about 44 degrees. Eye movements were recorded at 240 Hz using a commercial videographic eye tracker. Recordings were made for four

subjects (one recording each) on the first sequence and six subjects (three recordings each) on the second sequence. In all cases, we show averages of the results for all subjects.

We extracted saccades from the eye movements made by the test subjects and, for each saccade, computed the distance between the saccade target and the closest salient location. To obtain these locations, the L most salient locations over all levels of the spatio-temporal pyramid were extracted from the most recent video frame prior to the start of the saccade. Of course, there is a certain delay between a stimulus and the saccade induced by it, and our model takes this into account through the temporal filtering in the pyramid.

Figure 1 plots the average squared error (distance between saccade target and closest salient location), normalized by dividing by the average squared saccade length. We plot this error for various values of L. As an aid to assessing the salient locations generated by the saliency measures H, S and K, we also plot the errors obtained using locations distributed randomly over the image. The parameters in the salient-location extraction algorithm were set to $\sigma = 35$ pixels and $W = 100$ pixels.

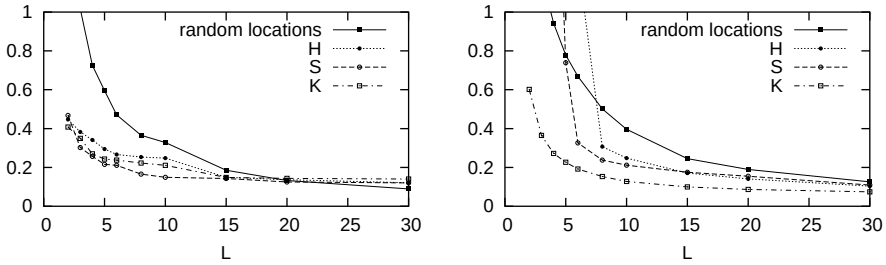

Figure 1: Average relative error for the H, S and K saliency measures and random locations on the synthetic sequence (left) and traffic scene (right). The horizontal axis plots L, the number of salient locations.

We observe that for small L, the three saliency measures perform significantly better than random locations. With increasing L, this difference decreases because the random locations begin to achieve a good coverage of the whole image, making it likely that one of the random locations will lie close to the saccade target.

Comparing the three saliency measures H, S and K, we note that H tends to produce the worst salient locations. On the synthetic sequence, S performs best, whereas on the traffic sequence, K performs best. Note in particular that on the traffic sequence the H and S measures start out with a very high error, becoming better than the random locations for an L of 5 to 8, whereas the K locations are better than the random locations from the start. This is because the sequence contains some high-frequency static content (a line of trees), to which the H and S measures assign the highest saliency. The moving cars, which the majority of saccades go to, only have salient locations generated for them after several locations have been generated for the trees. This explains the sharp drop in both curves when the first salient locations are generated for the moving cars. The K measure, in contrast, is more sensitive to motion and assigns the highest saliency to the moving cars. This explains the good performance of the K measure from the start. Based on these results, we conclude that the S and K measures yield good salient locations.

4 Conclusions and Outlook

We have presented three saliency measures and a salient location extraction algorithm based on the mechanism of selective lateral inhibition and validated them against locations actually attended by observers watching test videos. Our results show that the model generates good predictions of where an observer will look in a dynamic scene.

Of the three structure-tensor-based saliency measures presented (H, S and K), we have shown that S and K give better results than the H measure. This confirms that the higher the intrinsic dimension, the higher the saliency. To obtain a definite result on the relative quality of the S and K measures, we intend to perform more experiments on a greater number of higher-resolution video sequences.

The next step then is to use the saliency information to suppress or enhance features in the image in order to guide the user's gaze along a given path. We believe that this type of gaze-contingent interactive display will bring a new quality to human-machine interaction and visual communication.

Acknowledgements

Research is supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01B. We thank SensoMotoric Instruments GmbH for eye-tracking support; data were obtained using their iViewX system.

References

[BDM03] Barth, E., Drewes, J., and Martinetz, T.: Individual predictions of eye-movements with dynamic scenes. In: Rogowitz, B. and Pappas, T. (Eds.), *Electronic Imaging 2003*. volume 5007. SPIE. 2003.

[BKBM] Böhme, M., Krause, C., Barth, E., and Martinetz, T.: Eye movement predictions enhanced by saccade detection. In: *Brain Inspired Cognitive Systems 2004*.

[BMS02] Boccignone, G., Marcelli, A., and Somma, G.: Analysis of dynamic scenes based on visual attention. In: *Proceedings of AIIA 2002*. Siena, Italy. 2002.

[IK01] Itti, L. and Koch, C.: Computational modelling of visual attention. *Nature Reviews Neuroscience*. 2(3):194–203. 2001.

[Ita] Information technology for active perception (Itap). `http://www.inb.uni-luebeck.de/Itap`.

[Org] The organic computing page. `http://www.organic-computing.org`.