SALIENCY-BASED MULTI-FOVEATED MPEG COMPRESSION

Nitin Dhavale and Laurent Itti

Computer Science Department - University of Southern California
Los Angeles, California, 90089-2520 - USA - *http://iLab.usc.edu*

ABSTRACT

Most current foveation strategies are limited to foveating sequences based on a direct measurement or an implicit assumption of the gaze direction. Such approaches often fail in unconstrained environments or when necessary equipment is absent. Alternatively, a computational model of visual attention may be used to predict visually salient locations. We describe such a neurobiological model of attention and its specific application to foveated video compression. The algorithm is demonstrated to be successful in foveating to Regions Of human Interest in a variety of video segments, including synthetic as well as natural scenes, and also gives good compression ratios.

1. INTRODUCTION

One of the characteristics of humans is the ability to direct their visual attention to specific objects (or locations) of interest in a scene. Attention determines what people see [16]. Naturally, the development of foveated vision systems depends on the ability to automatically predict the locations of regions of human interest, and to foveate onto those regions. Several different approaches to predict areas of maximal human interest in a scene have been proposed. Privitera & Stark [13] use a computational algorithm based on experimentally-determined scanpath data. Their idea is to apply a list of Image Processing Algorithms (IPAs) to get a sequence of algorithmically-defined regions of interest (aROIs). The experimentally obtained human regions of interest (hROIs) are compared to the aROIs by analyzing the scanpaths arising from each IPA. A subgroup of the algorithms that maximizes the similarity between the aROIs and hROIs is selected and used to predict scanpaths. Moghaddamand & Pentland [7] examine fixation selection based on low level feature selection like contrast, edges, object similarity. Doll *et al.* [19] draw from previous vision research to predict a viewer's ability to discriminate pattern and color differences. A probability of fixation is estimated for each object in the scene based on the object's contrast, color, motion, and similarity to both the target and background. A special feature of their implementation is a signal detection theory routine to handle trade-offs between detections and false alarms. Other algorithms that are tuned to detect specific objects like faces or human figures have also been proposed.

Here we address two fundamental limitations of these previous approaches: First, we use a neurobiological model of attention [6] to select regions of interest in a manner that is fully automatic yet yields good agreement with human eye movement data in unconstrained environments [11], while most previous approaches have been limited to constrained environments. Second, we extend the model such as to not only yield one ROI in each frame, but possibly several regions *(multi-foveation)* or even a continuous, graded measure of interest. This measure is used to increasingly degrade (blur) the image far away from ROIs, such as to increase overall compression ratio. The entire system is tested on a variety of outdoors, indoors and synthetic (game console) MPEG-1 video clips, and yields an additional compression factor of 2-3 with degradations that are hardly noticeable by human observers.

2. DETERMINATION OF FIXATION LOCATIONS

Low-level features are extracted from each input image using a set of linear filters tuned to specific features like color, motion, orientations and intensity, like in [6]. This decomposition is performed at nine spatial scales using Gaussian pyramids. The output from the 72 channels is then combined into a unique saliency map. The saliency map is fed to a Winner-Take-All to find the locations of a fixed number of perceptually salient objects in the scene **(Fig. 1)**. This is motivated by recent experiments suggesting that subjects can allocate attention to 4-5 objects at the same time [14].

3. FOVEATION AND MPEG ENCODING

It is a challenging task to attempt to determine the single most interesting location in a scene; indeed, human visual attention is affected by factors such as culture, age, task at hand and psychological state of the observer [16], so that there is no good unique solution to the problem of finding where an average observer would look. We aim to make the quality perceived by the wide majority of the participants

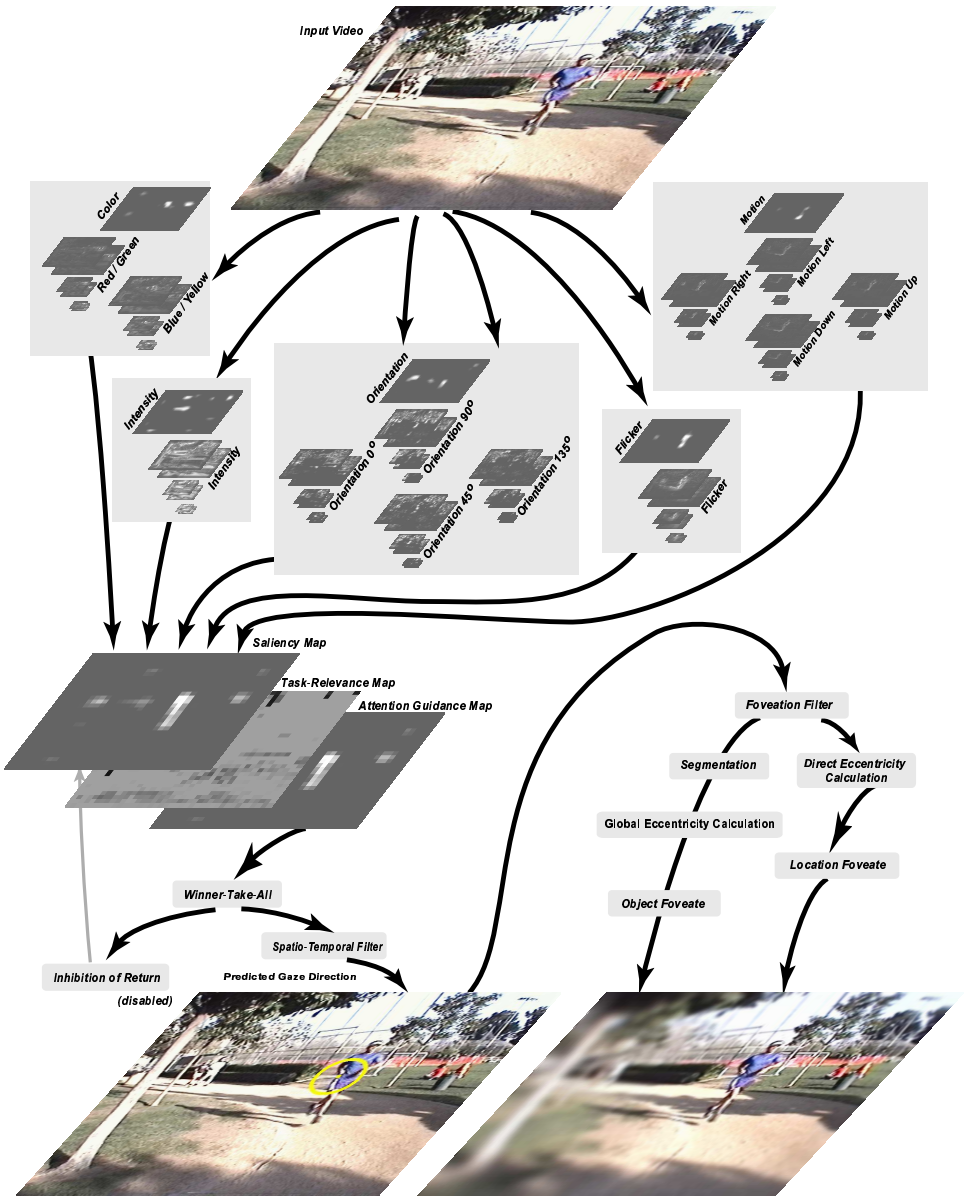

Figure 1: Overview of the model.

higher. Our first approach to this end is a multi-foveated approach, in which more than one region of interest is computed on each frame.

Currently, there is no unanimity as to whether attention selects perceptual objects [5, 10, 18] or it just acts like a spotlight that illuminates locations in space [12, 8]. So we implement both object-based and location-based foveation.

The spatial frequency response for any region in the retina can be modeled by scaling the foveal response with a factor based on retinal eccentricity e [17], such that [15]:

$$f_e = \begin{cases} \frac{M_0}{(0.3e+1)^2\sigma} & \text{if } |e| > 5.79 \\ 1.0 & \text{otherwise} \end{cases} \quad (1)$$

where f_e is the threshold frequency at eccentricity e, M_o is the highest spatial frequency that can be resolved at the fovea, and σ is the blur factor (larger values yield larger blurs). The value of M_o is simply $60c/^\circ$ [4]

Also, the highest frequency f_d that can be represented faithfully by a digital monitor is:

$$f_d = \frac{\pi N d}{360} \quad (2)$$

where N is the width of the screen in pixels and d is the viewing distance.

Combining (1) with (2) we obtain:

$$f_m = \min(f_e, f_d) \tag{3}$$

By the sampling theorem, the smallest size pixel that can be resolved at an eccentricity e is given by:

$$s = \frac{0.5}{f_m} \tag{4}$$

For the location-based foveation scheme, the eccentricity for any particular point is calculated using:

$$e = \tan^{-1} \frac{\|X - X^s\|}{Nd} \tag{5}$$

where $X = (x, y)$ and $X^s = (x_s, y_s)$ are the coordinates of the most salient location. In the more complex case of object-based foveation, the eccentricity value for a pixel at position $P(x, y)$ can be calculated using the chamfer transform of [2] or the exact Euclidean distance transform due to [1]. If the shortest distance between point $P(x, y)$ and an object is $g(x, y)$, the eccentricity is calculated using:

$$e = \tan^{-1} \frac{g(x, y)}{Nd} \tag{6}$$

In the case of foveation with a single ROI, the size of each pixel in the final image is calculated using equation (4). The pixel is then smoothed in a circle of radius $s/2$ by taking an average of all the pixels with a Gaussian weight:

$$V(x, y) = \sum_{i=-s/2}^{i \leq s/2} \sum_{j=-s/2}^{j \leq s/2} v(x + i, y + j) e^{-\frac{(x+i)^2 + (y+j)^2}{2s^2}} \tag{7}$$

where $V(x, y)$ is the resulting pixel value at (x, y), and $v(x, y)$ is the source image pixel value at (x, y).

A fast implementation of eccentricity-dependent blur can also be obtained using a Gaussian pyramid [3], which achieves foveation by interpolating across the levels of a Gaussian pyramid.

If there are n foveas, then there are effectively n foveated output images. The pixel value in the final foveated image is then given by:

$$S(x, y) = \max_{i \in [1, \ldots, n]} (w_i V_i(x, y)) \tag{8}$$

where w_i is the saliency value at the i^{th} location normalized to the maximum value, $V_i(x, y)$ is the pixel value in the i^{th} foveated image at (x, y) calculated using (7) or by using a Gaussian pyramid, and $S(x, y)$ is the pixel value in the final foveated image at position (x, y).

However, this calculation in quite heavy on the resources. Instead, one can use:

$$k = \min_{i \in [1, \ldots, n]} (w_i g_i(x, y)) \tag{9}$$

$$S(x, y) = V_k(x, y) \tag{10}$$

where the symbols are as before.

One of the side effects of multi-foveation is to smooth out the 'beauty jumps.' Frames are temporally averaged so that the foveal shifts appear smooth. A binomially weighted average of the distance maps due to frames in a cache buffer and a look-ahead buffer is used in calculating the eccentricity and applying the foveation filter:

$$T_i(x, y) = \frac{\sum_{j=i-l}^{i+r} S_j(x, y) w_j}{\sum_{j=i-l}^{i+r} w_j} \tag{11}$$

where T_i is the i^{th} output frame, S_j is the j^{th} foveated frame, W_j is the weight of the j^{th} frame and l and r are the widths of left and right ends of the cache window measured from the current output frame.

4. RESULTS

The algorithm has been tested with a variety of indoors, outdoors and synthetic video clips (captured from game console outputs). Overall, the regions of interest picked by the algorithm make sense to human observers, to the point that it is often difficult to notice the parafoveal blur, as one typically fixates to one of the most salient locations computed by the algorithm.

Fig. 2 shows the compression ratios obtained using our algorithm on a single movie clip. Each graph was plotted keeping the number of foveation points constant. Compression ratios of about 1.8 could be achieved without significant deterioration of quality as perceived by normal observers. As can be seen, compression for the object-based foveation were lower than those obtained for the location-based algorithm.

5. CONCLUSION AND FUTURE WORK

Although our algorithm works well with most scenes tested so far, it is not without limitations. For one, the algorithm does not take into account the fact that brain computations are in object-centered frame of reference. Second, we are only beginning to include task-oriented top-down influences into our model. Our algorithm would have to include object recognition and task-based biasing, as proposed in [9].

Further testing and validation of the algorithm will involve recording of eye movements from human observers

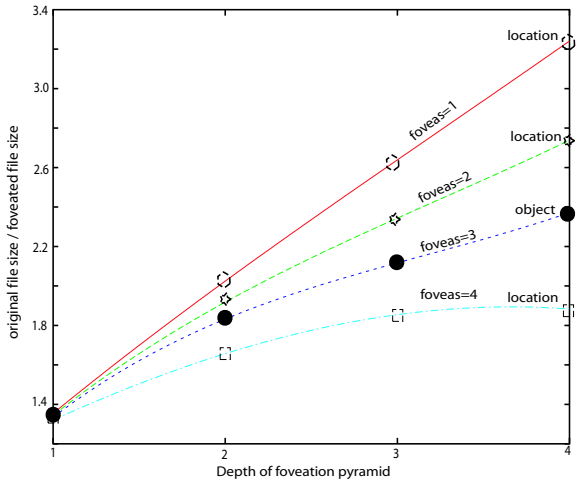

Fig 2(a): Compression ratios for location/object based foveation algorithms without averaging.

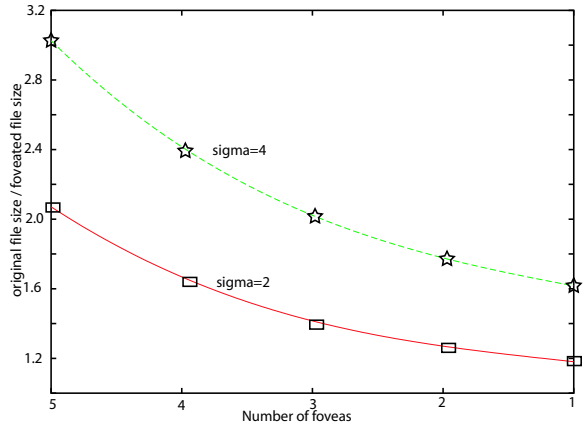

Fig 2(b): Compression ratios for location based foveation with averaging.

Figure 2: Experimental results

watching the video clips of interest. A particularly interesting issue in this context will be to determine whether our algorithm may be used in an iterative closed loop aimed at determining the optimal amount of blur that may be tolerated: Indeed, too much extra-foveal degradation should yield artifacts that would be perceived as salient by observers, and thus should also be picked up by the model if run a second time, on the foveated clip.

Acknowledgements: This work is supported by NSF, NEI, NIMA and the Zumberge Fund.

6. REFERENCES

[1] J.B.T.M. Roerdink A. Meijster and W.H. Hesselink. A general algorithm for computing distance transforms in linear time. *Mathematical Morphology and its Applications to Image and Signal Processing, J. Goutsias, L. Vincent and D. S. Bloomberg (eds.), Kluwer*, pages 331–340, 2000.

[2] G Borgefors. Distance transformations in digital images. In *CVGIP: Image Understanding*, volume 54, page 301, 1991.

[3] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31,4:532–540, 1983.

[4] F. W. Campbell and R. W. Gubisch. Optical quality of the human eye. *J Physiol*, 186(3):558–578, Oct 1966.

[5] P. Halligan G. Fink, R. Dolan and C. Frith. Space-based and object-based visual attention: Shared and specifi c neural domains. *Brain*, pages 2013–2028, 1997.

[6] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res*, 40(10-12):1489–1506, 2000.

[7] B. Moghaddamand and A. Pentland. probabilistic visual learning for object detection. *Proc of fifth conf. on computer vision*, pages 786–793, 1995.

[8] K. Cave N. Bichot and H. Pashler. Visual selection mediated by location: Feautre based selection of non-contiguos locations. *Perception and Psychophysics*, pages 403–423, 1999.

[9] V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany*, pages 453–461, Nov 2002.

[10] U. Neisser and R. Becklen. Selective looking: Attending to visually specifi ed events. *Cognitive Psychology*, pages 484–494, 1975.

[11] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Res*, 42(1):107–123, Jan 2002.

[12] M. Posner and Y. Cohen. Components of visual orienting. *Attention and Performance*, pages 55–66, 1984.

[13] C. Privitera and L. Stark. Algorithms for defi ning visual regions-of-interest: Comparision with eye fi xations. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 22:970–982, 1999.

[14] Z. Pylyshyn and R. Storm. Tracking multiple independent targets: Evidence for parallel tracking mechanism. *Spatial Vision*, 3:179–197, 1988.

[15] M. Reddy. Robust, sensor-independent target detection and recognition based on computational models of human vision. *PhD thesis, University of Edinburgh*, 1997.

[16] R. A. Rensink. Change detection. *Annu Rev Psychol*, 53:245–277, 2002.

[17] J. Rovamo and V. Virsu. An estimation and application of the human cortical magnifi cation factor. *Exp Brain Res*, 37(3):495–510, 1979.

[18] D. Simons and C. Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 2002.

[19] A.A. Wasilewski T.J. Doll, S.W. McWhorter and D.E. Schmieder. Robust, sensor-independent target detection and recognition based on computational models of human vision. *Optical Engineering*, 1998.