MULTI-FOVEATION FILTERING

T. Popkin, A. Cavallaro

Multimedia and Vision Group,
Queen Mary, University of London, E1 4NS, UK

D. Hands

British Telecommunications PLC,
Martlesham Heath, IP5 3RE, UK

ABSTRACT

We present a method for computing a function of average multi-viewer eye sensitivity based on the Geisler & Perry contrast threshold formula, and, from this, the cut-off frequency map (as used in *foveation filtering*) that is optimal in the sense of discarding frequencies in least-noticeable-first order. Existing approaches usually solve the multi-viewer foveation problem as a number of single-viewer foveations, effectively taking collective sensitivity to be the maximum of the individual viewer eye sensitivities. This has inherent problems such as over-sensitivity to outliers which are not problems with the proposed approach. Furthermore, the proposed approach can be employed in the infinite-viewer (probability-based) scenario without additional cost.

Index Terms— Contrast sensitivity, foveated image processing, foveation, foveation filtering, multiresolution.

1. INTRODUCTION

Lossy image and video coding techniques which aim to exploit the spatially-variant acuity of the eye by directly matching this acuity are known as *foveated* coding techniques. These have received an extensive amount of research (see [1] for a review), motivated by the stark difference in resolution between the point of fixation and the periphery of the retina. Their benefits have been particularly demonstrated in *gaze-contingent* (single, known fixation point) coding, by results such as an 18.8-to-1 reduction in bandwidth with minimal perceived loss of quality [2]. Gaze-contingent coding is suited to scenarios in which real-time eye tracking is available. However, in typical coding scenarios no eye tracking is available and there may be any number of viewers, gazing at different points. Therefore, a number attempts have been made [1, 3, 4] to extend foveated coding to the multi-viewer scenario, or ideally to the probability-based (infinite-viewer) scenario in which a *saliency map* may be sourced from an *attention model*, and most of these attempts effectively assume the multi-viewer sensitivity at each location to be the maximum of the individual viewer senstivities.

This work was jointly supported by the Engineering and Physical Sciences Research Council (UK) and British Telecommunications PLC.

We propose herein an alternative definition of a multi-viewer or saliency-based sensitivity function which aims to satisfy a collective average viewer sensitivity. We propose an efficient method for the computation of this function and for computing the corresponding map of cut-off frequencies for a given value of the overall sensitivity parameter, and we compare it with an existing multi-viewer cut-off map generation technique.

Further discussion of the related background is given in Sec. 2. The model is defined in Sec. 3. The algorithm for computing a single sensitivity value is given in Sec. 4 and the algorithm for computing a whole cut-off map is given in Sec. 5. Experimental results and comparisons are discussed in Sec. 6. In Sec. 7 we conclude the paper.

2. BACKGROUND

Foveated coding techniques usually assume a simplified, radially-symmetric model of the spatial variation of acuity. A popular model is the *contrast threshold* formula of Geisler & Perry [5], from which *contrast sensitivity* can be defined as the reciprocal of the contrast threshold [6]:

$$CS(f,e) = \frac{1}{CT(f,e)} = \frac{1}{CT_0} \exp\Big(-\alpha f \frac{e+e_2}{e_2}\Big), \qquad (1)$$

for each eccentricity e (deg) and spatial frequency f (cycles/deg), with constants $e_2 = 2.3$, $\alpha = 0.106$ and $CT_0 = 1/64$. Assuming a strict maximum of 1 for the contrast threshold (that is, setting $CT(f,e) = 1$) gives a cut-off frequency $f_c(e) = e_2 \ln(1/CT_0)/((e+e_2)\alpha)$ [6]. This allows a spatial map of cut-off frequencies (that is, blur levels) to be constructed, which can be employed in *foveation filtering* or in a DCT coefficient cut-off scheme within an image or video encoder. Extending this to the multi-viewer scenario, the majority of existing approaches can be regarded as aiming for distortion that lies below the contrast threshold of every viewer [1, 4, 7, 8]. These approaches work as if the collective sensitivity of each frequency at each location should be the maximum of the sensitivities of individual viewers. Because of the radial symmetry of the sensitivity function, this approach is equivalent to spatially partitioning the multi-viewer problem into a number of single-viewer foveations around the nearest fixation points, from which the inverse (i.e., the

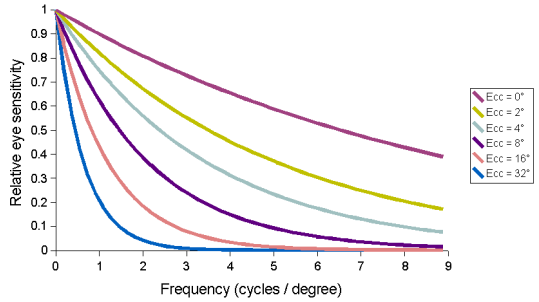

Fig. 1. The assumed eye sensitivity model at different eccentricities.

cut-off frequency) can be easily computed. However, this is not ideal because, for example, any number of co-fixated viewers are treated exactly as a single viewer with the given fixation point, and local fixation point density is disregarded. Furthermore, within regions which attract fixation, as the number of viewers becomes large and the inter-fixation-point distance becomes small, the solution locally converges within these regions to that of ordinary, spatially-uniform encoding, therefore losing some or all of the coding advantages of having knowledge of human fixation. We overcome these problems in Sec. 3 by employing an additive combination of sensitivities.

3. MULTI-VIEWER SENSITIVITY MODEL

In this section we define the multi-viewer sensitivity function that we employ.

Eq. (1) gives a measure of the sensitivity of a human eye to a given frequency component in a given direction. From this, we derive a normalised sensitivity function $s:\mathbb{R}_+\times\mathbb{R}_+\rightarrow\mathbb{R}$, defined as follows: $s(f,e) = \exp(-(e+e_2)\,\alpha f/e_2)$ for all $e,f\in\mathbb{R}_+$, with constants α and e_2 as specified earlier. $\mathbb{R}_+$ herein represents the set $[0,\infty)$; that is, all non-negative real numbers. This function is illustrated in Fig. 1.

To convert this to the image domain, we assume that the viewer is positioned so that he has head-on viewing of the fixation point. Therefore, given that the fixation point is located at $\mathbf{y}$ and the viewing distance is d (in pixels), as an approximation (neglecting trigonometry), the eccentricity of image location $\mathbf{x}$ will be $360\|\mathbf{x}-\mathbf{y}\|/2\pi d$. Now, define function $a:\mathbb{R}_+\rightarrow\mathbb{R}_+$ as follows:

$$a(r) = (360r/2\pi d \;+\; e_2)\,\alpha/e_2 \tag{2}$$

for all r. Let the set $D=\{0,...,W\text{–}1\}\times\{0,...,H\text{–}1\}$ represent the domain of any $W\times H$ image. Then, a sensitivity function $s_{\mathbf{y}}:D\times\mathbb{R}_+\rightarrow[0,1]$, for a given fixation point $\mathbf{y}\in D$, can be defined as $s_{\mathbf{y}}(\mathbf{x},f) = \exp(-a(\|\mathbf{x}-\mathbf{y}\|)\,f)$ for all $\mathbf{x}$ and f.

We extend this to the multi-viewer and infinite-viewer scenarios by summation. In the infinite-viewer scenario, in which we have a fixation probability density map (saliency map) $\mu:D\rightarrow[0,1]$, the infinite-viewer sensitivity level $S_{\mu,\mathbf{x}}(f)$ for location $\mathbf{x}$ and frequency f is

$$S_{\mu,\mathbf{x}}(f) = \textstyle\sum_{\mathbf{y}\in D}\,\mu(\mathbf{y})\exp(-a(\|\mathbf{x}-\mathbf{y}\|)f). \tag{3}$$

Note that the interpretation of a finite-viewer sensitivity function from this infinite-viewer function can be performed by setting map μ to be a sum of 2-dimensional Dirac delta functions.

4. COMPUTING MULTI-VIEWER SENSITIVITY

To compute each $S_{\mu,\mathbf{x}}(f)$ value by interpreting Eq. (3) verbatim would be prohibitive, as each sensitivity value would involve a sum of HW terms. However, a close approximation of this computation can be performed using a faster approach which we will now describe.

Define a family $E = \{e_\beta : \beta\in[0,A]\}$ of functions $e_\beta:[0,F]\rightarrow\mathbb{R}$, where $A = a(\sqrt{(H-1)^2+(W-1)^2})$ is the maximum $a(\|\mathbf{x}-\mathbf{y}\|)$ value that can occur in Eq. (3) and $e_\beta(f) = \exp(-\beta f)$ for all $f\in[0,F]$. Here, F is the maximum-representable frequency, which for head-on viewing equates to the pixel-diagonal Nyquist frequency; that is, the maximum-representable number of cycles per $\sqrt{2}$ pixel widths, i.e. $\sqrt{1/2}$ cycles per pixel, i.e. $\sqrt{1/2}\times 2\pi d/360$ cycles per degree, where d is the viewing distance, as before. Consider a function $z:[0,F]\times[0,F]\rightarrow\mathbb{R}$, defined such that

$$z(f_1,f_2) = \int_0^A e_\beta(f_1)e_\beta(f_2)d\beta \tag{4}$$

for all f_1 and f_2. Now, when z is approximated by a discrete-domain matrix Z, this matrix is symmetric, and hence can by diagonalised by an orthonormal basis of eigenvectors [9, p.379]. This process can be used to approximate the orthonormal eigenfunctions $b_1,b_2,b_3,...,$ of z. It transpires that all except a small number of the eigenvalues of z are very close to zero, the result being that each function e_β can be approximated closely by a linear combination of the first N principal eigenfunctions, $b_1,...,b_N$, for a suitably chosen N. That is for every β and f,

$$e_\beta(f) = \exp(-\beta f) \approx \textstyle\sum_{n=1}^{N} c_n(\beta)b_n(f), \tag{5}$$

where each function $c_n:[0,A]\rightarrow\mathbb{R}$ is defined as

$$c_n(\beta) = \int_0^F b_n(f)e_\beta(f)df \tag{6}$$

for all β; that is, thinking in vector terms, each scalar $c_n(\beta)\in\mathbb{R}$ is the component of vector e_β in the direction of vector b_n. For practicality, we approximate each $c_n(\beta)$ using a discrete-domain summation, and store them in a lookup table for use thereafter in all approximations of the e_β functions.

These linear combinations of the N chosen eigenfunctions can be used to approximate the $\exp(-a(\|\mathbf{x}-\mathbf{y}\|)f)$

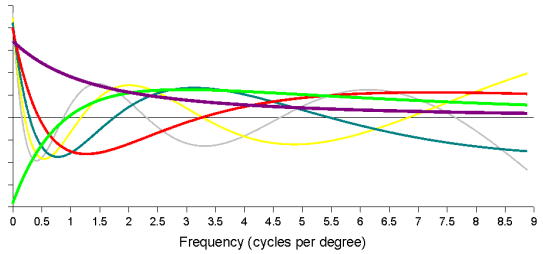

Fig. 2. The basis functions used in the computation of the sensitivity-v-frequency curves.

part of Eq. (3), and hence they form an approximate basis for the space of possible sensitivity-versus-frequency curves. We have found empirically that with $N=6$, $H=240$, $W=360$ and $d=3H$, the worst root mean squared error of any e_β approximation is roughly 0.0001. These first six eigenfunctions are depicted in Fig. 2.

Substituting Eq. (5), with $\beta = a(\|\mathbf{x}-\mathbf{y}\|)$, into Eq. (3), gives a sum of sums, $S_{\mu,\mathbf{x}}(f) \approx \sum_{\mathbf{y}\in D} \mu(\mathbf{y}) \sum_{n=1}^{N} c_n(a(\|\mathbf{x}-\mathbf{y}\|))\, b_n(f)$. This can be written as

$$S_{\mu,\mathbf{x}}(f) \approx \sum_{\mathbf{y}\in D} \mu(\mathbf{y}) \sum_{n=1}^{N} C_n(\mathbf{x}-\mathbf{y})\, b_n(f),$$

which can be rearranged as a sum of convolutions, $S_{\mu,\mathbf{x}}(f) \approx \sum_{n=1}^{N} b_n(f)\, (\mu * C_n)(\mathbf{x})$ where each 2-dimensional-domained function $C_n : \breve{D} \to \mathbb{R}$ is defined in terms of the corresponding 1-dimensional-domained function c_n as

$$C_n(\mathbf{w}) = c_n(-a(\|\mathbf{w}\|)), \tag{7}$$

for all $\mathbf{w}$, where $\breve{D} = \{\mathbf{x}-\mathbf{y} : \mathbf{x}, \mathbf{y} \in D\}$ is an extended version of the image domain D. Therefore,

$$S_{\mu,\mathbf{x}}(f) \approx \sum_{n=1}^{N} b_n(f)\, \psi_{\mu,n}(\mathbf{x}), \tag{8}$$

where each coefficient map $\psi_{\mu,n} : D \to \mathbb{R}$ is defined as a convolution $\psi_{\mu,n} = \mu * C_n$.

Each of these N convolutions, which can be performed using a fast convolution technique [9, p. 449], needs only to be done once for each saliency map μ and thereafter stored in a look-up table, after which the same $\psi_{\mu,n}$ maps will be looked up and used for the computation of each sensitivity value $S_{\mu,\mathbf{x}}(f)$ for a given location $\mathbf{x}$ and frequency f. Note also that each eigenfunction b_n and map C_n only need to be computed once for given values of W, H and d, and thereafter re-used for all $W \times H$ saliency maps without any need for recomputation. Applying Eq. (8) for a single sensitivity f and location $\mathbf{x}$ will cost only N look-up operations (one for each b_n), N multiplications and $N-1$ additions, so is an order $\mathcal{O}(N)$ operation. However, the dominant part of the computation will be the computation of the maps $\psi_{\mu,1}, ..., \psi_{\mu,N}$,

Fig. 3. Examples of cut-off maps produced by the proposed approach with blur levels 10%, 30%, 50%, 70% and 90%. (White = max. frequency; black = zero.)

each of which will be of order $\mathcal{O}(HW \log(HW))$, but which only need to be done once for each saliency map μ. Therefore, the cost per pixel is of order $\mathcal{O}(N \log(HW))$, which can be regarded as $\mathcal{O}(\log(HW))$ as N is fixed (we used $N=6$). This compares with $\mathcal{O}(HW)$ per pixel for a verbatim implementation of Eq. (3).

The next section explains how to create a map of cut-off frequencies.

5. COMPUTING A CUT-OFF FREQUENCY MAP

Consider the solution f to the equation $\gamma = S_{\mu,\mathbf{x}}(f)$ for a given location $\mathbf{x} \in D$, satisfying some given overall sensitivity level γ, and where $S_{\mu,\mathbf{x}}(f)$ is computed using Eq. (8). If this were solved for every $\mathbf{x} \in D$, the result would be a spatial map of frequencies with this given sensitivity. Combining this with the knowledge that each $S_{\mu,\mathbf{x}}$ is a strictly decreasing function, this map can be interpreted as a spatial map $\phi_{\mu,\gamma}$ of the lowest frequencies that have lower sensitivity than γ (that is, a cut-off map) defined as follows: $\phi_{\mu,\gamma}(\mathbf{x}) = S^{-1}_{\mu,\mathbf{x}}(\gamma)$ for all $\mathbf{x} \in D$. Thus, for lossy coding purposes, each cut-off map $\phi_{\mu,\gamma}$ is optimal in the sense of discarding least-noticeable information first. Using our forward computation for function $S_{\mu,\mathbf{x}}$, we compute the inverse $S^{-1}_{\mu,\mathbf{x}}$ by a binary search (*bisection* [9, p. 277]).

Fig. 3 present examples of cut-off frequency maps from the proposed technique, with varying levels of the quantization parameter. Further examples are shown in the next section.

6. RESULTS AND COMPARISON

We compare the proposed technique with the multi-viewer cut-off map generation scheme by Sheikh *et al.* [4], which is an example of a technique that takes collective sensitivity to be the maximum of individual sensitivities, and which has the additional characteristic that its overall blurring can be controlled by a tuning parameter. The output of the proposed technique and two variants of the Sheikh *et al.* are shown in Fig. 4. In the first variant, we used their suggested parameters (block width 16, 8 quantization levels). The second was a continuous variant of their technique (block width 1, unlimited quantization levels), in which, for consistency with the proposed approach, we increased their hard-limited maximum frequency from 0.5 cyc/pixel up to the pixel-diagonal

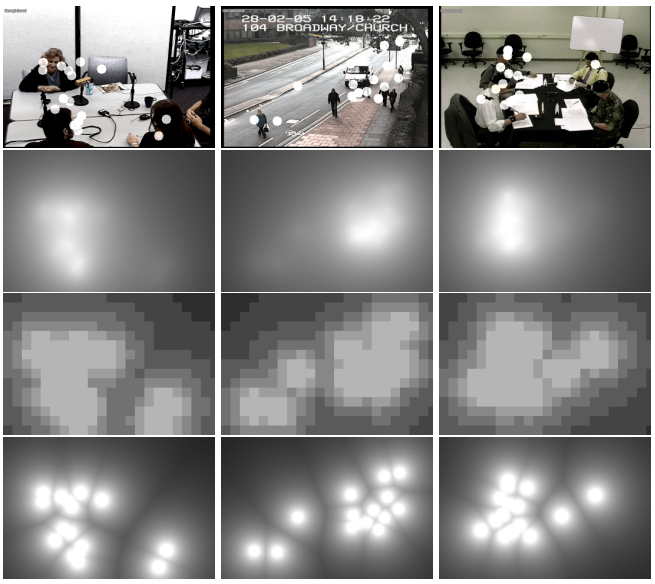

Fig. 4. Cut-off frequency maps with blurring level 60%. Top row: frames highlighted with fixation points. Second row: proposed method. Third row: [4] as published. Bottom row: continous variant of [4]. Left Columns: left: 169$^{\text{th}}$ frame of sequence *cmu_2_cam3_1*; middle: 4674$^{\text{th}}$ frame of *pvtra102a10*; right: 153$^{\text{rd}}$ frame of *vt_2_cam2_1*. Frames sourced from *CLEAR 2006* and *CLEAR 2007* video datasets.

Nyquist frequency of $\sqrt{0.5}$ cyc/pixel.

The cut-off maps were generated from fixation points collected using an eye tracker from 16 subjects independently viewing three video sequences (see Fig. 4). Throughout, we assumed a viewing distance of three times the frame height (i.e., a distance of 720 pixel widths). The cut-off sensitivity parameter, γ, which controls the general level of blurring, was chosen automatically using a binary search scheme to home-in on the value that produces a desired overall percentage of frequencies cut off, taking into account the relationship between frequency magnitude and the number of frequency bins of lower magnitude in 2-dimensional frequency space. We applied the same approach to the technique of Sheikh *et al.*.

The bottom row of Fig. 4 highlights the nature of the Sheikh *et al.* method as effectively partitioning video frame according to the nearest fixation point to each location, and separately computing a single-viewer cut-off map for each location.

A key issue in deciding which technique provides the more appropriate cut-off map is that of how they handle outlying fixation points. Fig. 4 shows the effect of a scenario with one or two fixation points outlying towards the right of the image. The proposed approach appears almost to neglect such points, with the main body of preserved frequencies surrounding the dominant cluster of fixations, whereas the Sheikh *et al.* technique effectively boosts the significance of the outliers, providing a more widely-spread region of higher resolution. Because the proposed approach concentrates the frequencies into a narrower area, the average resolution over this area will be higher. Also, the proposed approach has a smoother variation in frequency, thus reducing the chances that local variation in blur level may itself be observed as an artifact.

7. CONCLUSION

We have proposed a multi-viewer spatio-frequency sensitivity model, based on the Geisler & Perry contrast threshold formula, and a technique for computing a spatial map of local cut-off frequencies that is optimal for lossy coding purposes in the sense of discarding visually least-noticeable frequencies first. We have compared the results of the proposed technique with the output of that of Sheikh *et al.*. An important advantage of the proposed approach is the capability of handling the infinite-viewer (saliency-based) scenario with equal cost to the multi-viewer scenario. Future work includes extending the proposed approach to sensitivity functions other than that of Geisler & Perry, and finally, incorporating it into a foveated video coding system.

8. REFERENCES

[1] Z. Wang and A. C. Bovik, "Foveated image and video coding," in *Digital Video, Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds., chapter 14, pp. 431–457. CRC Press, 2006, ISBN 0-8247-2777-0.

[2] P. Kortum and W. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," *Proc. SPIE, Vol. 2657*, pp. 350–360, Apr. 1996.

[3] N. Dhavale and L. Itti, "Saliency-based multifoveated MPEG compression," in *Proc. 7th International Symposium on Signal Processing and Its Applications*, July 2003, vol. 1, pp. 229–232.

[4] H. R. Sheikh, B. L. Evans and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real-Time Imaging*, vol. 9, no. 1, pp. 27–40, Feb. 2003.

[5] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communincation," *Proc. SPIE, Vol. 3299*, pp. 294–305, July 1998.

[6] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

[7] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 911–992, July 2001.

[8] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 149–162, Feb. 2003.

[9] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes in Pascal*, Cambridge University Press, 1989, ISBN 0-521-37516-9.