Comparing image resamplers via a model of the human vision system

Richard Harvey, Stephen King, Richard Aldridge
and J. Andrew Bangham
School of Information Systems
University of East Anglia, Norwich NR4 7TJ
Tel: +44 1603 593257, Fax: +44 1603 593345.
{rwh,ssk,rva,ab}@sys.uea.ac.uk

Abstract

Interpolating images is a very common task but many interpolation filters are designed using error metrics that, in comparison to the known human vision system, are fairly crude. In this paper a model of the human vision system is used to produce a quality score that is useful for comparing the errors due to image resampling. It is tested with a number of image resamplers. We conclude that such a model produces a measure of image quality that is superior to signal-to-error ratio, but that further refinements are necessary.

1 Image resampling problems

In practical computer vision systems resampling is commonplace. Format conversion, rectification, and warping are all operations that involve resampling. Therefore there is considerable interest in developing resampling schemes that operate with any scaling factor without user intervention. Most resampling methods assume that the original intensity distribution is correctly sampled and construct new images by linear filtering followed by resampling. It is the design of the filter that is usually the subject for discussion since one seeks to remove the aliases due to sampling in some mathematically tractable fashion. If a quantitative quality measure is given for an interpolation technique it is usually obtained by contracting then expanding the image back to its original size and computing the error between it and the original. For example, in a recent paper [12], images are contracted and expanded by a factor $\sqrt{2}$ and compared quantitatively on the basis of signal-to-error ratios. However, several studies [15, 6] have concluded that both signal-to-error ratio (SER) [4], and gradient-weighted error [16], are unsatisfactory measures of perceived image similarity. In Figure 1, for example, images have been corrupted with additive noise, with different spatial correlation functions, but with identical signal-to-error ratios of 20 dB. At normal viewing distance readers with normal eyesight will find the left-hand image in Figure 1 considerably more disturbing than the right.

In the absence of quantitative models of the human vision system a crude approach based on signal-to-error ratios is understandable. However, recently, simple models of the human vision system have become available ([15, 18, 17] for example). In this paper we adapt such a model, originally developed for examining coding errors, and test its effectiveness with resampled images.

Figure 1: Images corrupted with additive Gaussian noise with a centre wavenumber of 8 π radians per degree (left) and 32 π radians per degree (right). Correct viewing distance is 51 cm.

2 Computing a perceptual score

The model [18] used here was designed for colour images and uses the opponent colour representation [13] but in this paper we restrict the discussion to monochrome images and use only the B/W channel (which is extremely close to the luminance (Y) channel). A schematic of the system is shown in Figure 2.

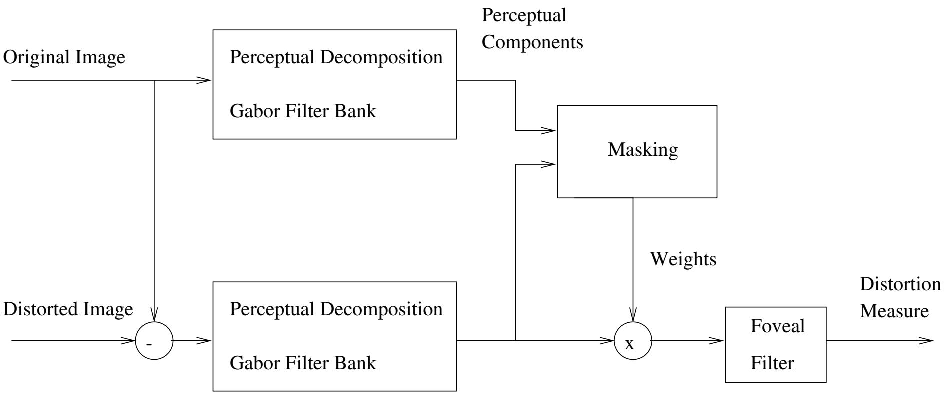

Figure 2: Schematic of the human vision model

Both the original image and the error are filtered into perceptual channels. The contrast of the original image is then evaluated and used to mask the error. This gives a distortion measure that is averaged in a manner that crudely models the fovea. The blocks labelled "Perceptual decomposition" consist of a set of Gabor filters. The first band-pass filter in the set is isotropic with zero response at wavenumber $\boldsymbol{k} = (k\cos\theta, k\sin\theta) = \boldsymbol{0}$ (to model insensitivity to global luminance level),

$$G(k) = e^{-\left(\frac{k}{\Delta k_a}\right)^2} - e^{-\left(\frac{k}{\Delta k_b}\right)^2}, \tag{1}$$

where k is the wavenumber measured in radians per degree of visual of angle and $\Delta k_a = 4\pi$ rad deg^{-1}, $\Delta k_b = \pi/2$ rad deg^{-1}. The other filters have a bandpass response centred

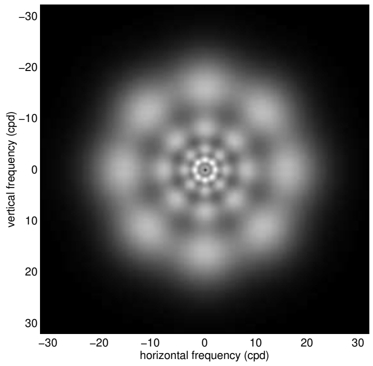

Figure 3: Response of the Gabor filter set in Fourier space. Axes are labelled in cycles per degree of visual angle.

on wavenumber $\boldsymbol{k}_0 = (k_0 \cos\theta_0, k_0 \sin\theta_0)$,

$$G(k_0, \theta_0) = e^{-\left(\frac{\log_2(k/k_0)}{\Delta k}\right)^2} e^{-\left(\frac{\theta-\theta_0}{\Delta\theta}\right)^2}, \tag{2}$$

where $\Delta k = 0.5$, $\Delta\theta = \pi/8$. The filters, shown in Figure 3, are chosen to model the visual channels [8]. Each channel of the distorted image is compared to the same channel from the original image and a masking model applied [5].

The masking model used here allows only within-channel masking and uses masking weights computed as the inverse the normalised detection threshold:

$$C_T = \begin{cases} 1 & C_M < C_{T_0} \\ \left(\frac{C_M}{C_{T_0}}\right)^{2.5} & C_M \geq C_{T_0} \end{cases} \tag{3}$$

where $C_{T_0} = (k_f C)^{0.62}$ is the detection threshold of the error in the absence of the masker. C is the error contrast and k_f is the contrast sensitivity function,

$$k_f = d\,(a + k_s/k_{s_0})\, e^{-(k_s/k_{s_0})^c}, \tag{4}$$

where $a = 0.0192, c = 1.1, d = 2.6$ and $k_{s_0} = 16\pi$ rad deg^{-1} are experimentally determined constants [9]. C_M is the contrast of the original image (the masker).

The masked error contrast is averaged using a disc shaped filter. The disc is chosen to subtend 2° so as to approximate the fovea. The final distortion is computed as

$$E(x, y) = \left(\frac{1}{N} \sum_{c=1}^{N} \left(\frac{1}{M} \sum_{x,y \in \boldsymbol{D}} |e(x, y)| \right)^4 \right)^{\frac{1}{4}} \tag{5}$$

where there are N channels, $\boldsymbol{D}$ is the set of M pixels in the foveal disc and $e(x, y)$ is the masked error signal at position x, y. The Minkowski sum in (5) is an attempt to weight errors in the same way as human observers [18]. $E(x, y)$ is called the Visual Difference Score.

Figure 4: Original 200 by 200 pixel test images (viewing distance 55 cm)

3 Results

Figures 5, 7 and 8 show the results of some putative image resampling techniques applied to the test images, shown in Figure 4. The first image contains black text (grey level 0) on a white background (grey level 255). Text rescaling via image resampling is interesting as character bitmaps are specially tuned to a particular resolution. The second image shows a simple landscape scene. The third image shows a human face and contains both over- and under-exposed regions. Each image in Figures 5, 7, and 8 has been contracted then expanded by a factor of four (not all of the methods allow expansion and contraction by anything other than integers). The methods that are compared here are:

1. Nearest neighbour interpolation;
2. Lowpass area m-sieving [1] to area ten with nearest neighbour resampling – the sieve is a recent extension of mathematical morphology;
3. Lowpass recursive one-dimensional sieving [2] (vertical then horizontal) with nearest neighbour resampling;
4. Bilinear interpolation;
5. Polynomial spline interpolation [12, 11] (here we use a first order B-spline);
6. FFT interpolation [3];
7. DCT interpolation [7];
8. Group mean reduction followed by an edge enhanced interpolation [16].

The resampled images were shown to a human viewing panel consisting of 13 people. Each person was asked to score image quality on a ten point scale. The images from all resampling methods were presented simultaneously. An uncorrupted image was also included as a control. Each observers scores were then ranked (the control image was always judged the best so was removed from the rankings) and the mean and standard deviation of rank computed across all observers.

The viewing panel identified method 7 as the best image (mean rank of 1.5 in Table 1). Method 7 also has a low visual difference score, but not the lowest. In fact signal-to-error ratio correctly identifies the image our panel thought least distorted.

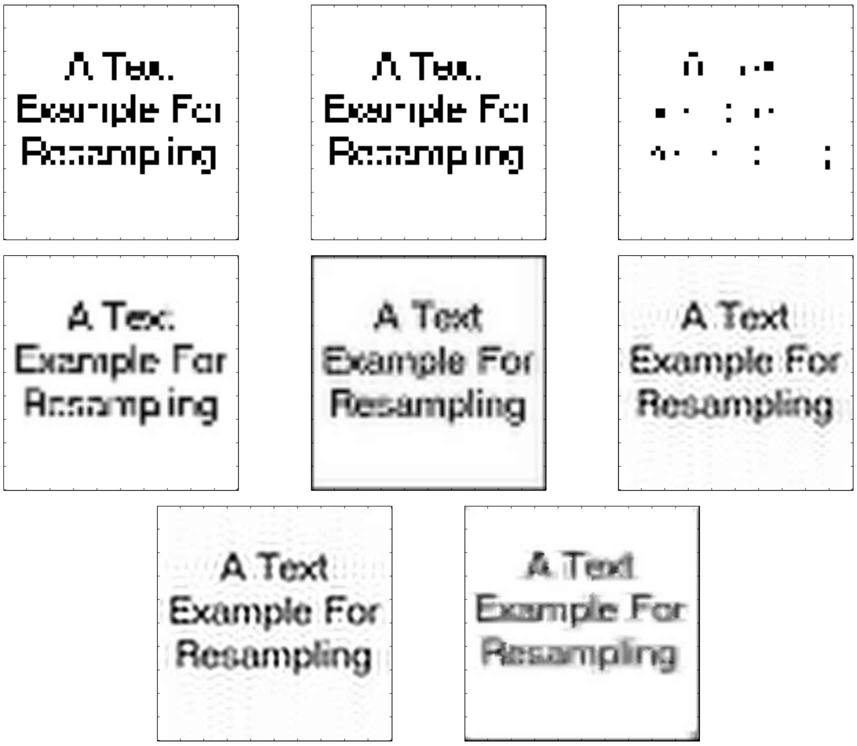

Figure 5: Text interpolated by, from left to right and top to bottom, methods 1 to 8. The correct viewing distance is 55 cm.

Method	Visual Difference Score		Signal-to-error	Human Ranking	
	Mode	Median		Mean	σ
1	1.00	0.76	11.7	4.4	1.0
2	0.96	0.74	11.7	4.5	1.1
3	0.20	0.25	10.4	5.7	0.8
4	0.06	0.20	13.6	2.9	0.9
5	0.08	0.06	14.0	2.2	1.1
6	0.14	0.20	14.3	2.1	0.6
7	0.12	0.18	14.7	1.5	0.7
8	0.09	0.14	12.2	3.0	1.2

Table 1: Scores for the text image of Figure 5.

Figure 5 shows the result of resampling the text image with these eight methods. Method 5 zero pads the image which gives a black border that we ignore here.

The visual difference score varies from pixel-to-pixel and its distribution may be complicated. For example in image 7 in Figure 5 the majority of pixels are in error, but the panel preferred this image to, say, image 1 in which the majority of pixels are uncorrupted. To some extent this effect is modelled by the foveal averaging, but clearly further investigation is necessary. Figure 6 shows histograms of the visual difference score after foveal averaging for three of the methods. The distributions are multimodal so a question arises. Do humans concentrate on the average error or some other distribution statistic? This question has not been resolved so we record the median and mode scores.

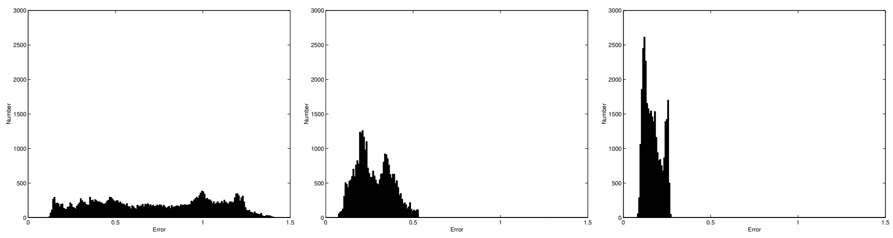

Figure 6: Histograms of Visual Difference Scores for Text images resampled by Methods 1 (left), 3 (centre) and 7 (right)

The vision model is adjusted so that one degree of visual arc is equivalent to 62 pixels (on our equipment this corresponds to a viewing distance of 1 m and puts the horizontal and vertical Nyquist frequencies at roughly 64π rad deg^{-1}). Viewing Figure 5 at 55 cm gives the same effect. Visual scores produced by modelling the human vision system are dependent on viewing distance [10] so the results presented here are only valid for the distances stated. Some care is needed when interpreting Figure 5 and Table 1 as the Visual Difference Scores are not meant to model how well we recognise text: they are a measure of image quality. For this reason previously unseen images such as those in Figure 7 may give more consistent scores.

Table 2 gives the scores for Figure 7. The panel of human observers chose methods 5, and 8 as having the lowest errors. Methods 5, and 8 also have low median and mode visual difference scores. Furthermore the median and mode are close together indicating a consistently low error over the whole image. Note that method 8 which is ranked highly by the panel and the visual difference scores has one of the lowest signal-to-error ratios, so in this case ranking by signal-to-error fails.

Figure 8 shows the face image. The scores are reported in Table 3. Again the panel preferred methods 5, and 8, and these had consistently low visual difference scores.

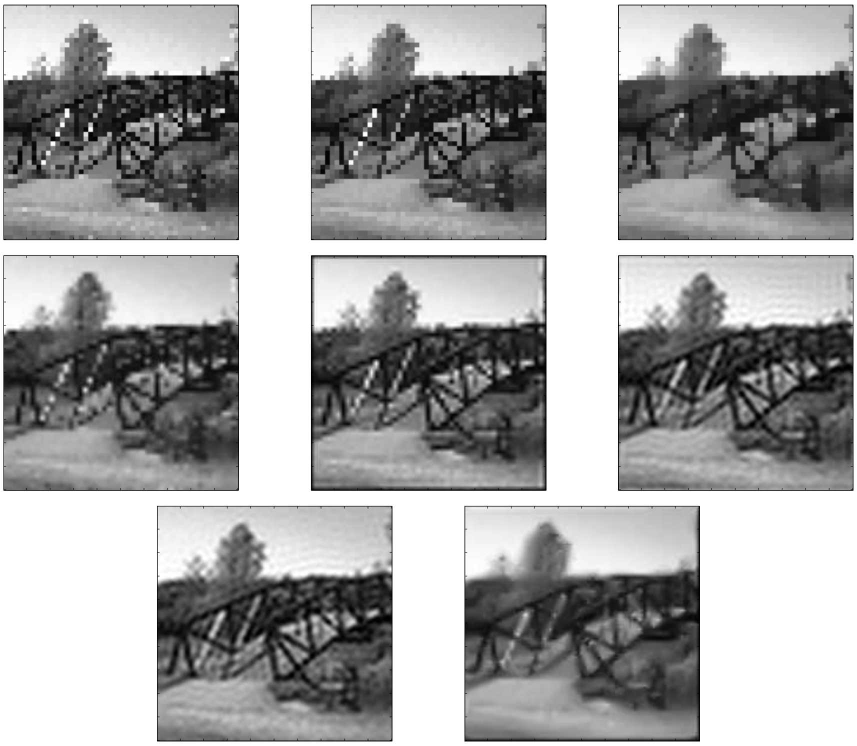

Figure 7: Bridge image interpolated by, from left to right and top to bottom, Methods 1 to 8. The correct viewing distance is 55 cm.

Method	Visual Difference Score		Signal-to-error	Human Ranking	
	Mode	Median		Mean	σ
1	0.30	0.79	17.6	3.3	1.3
2	0.23	0.50	18.1	4.1	1.1
3	0.17	0.20	18.2	4.6	0.9
4	0.07	0.27	19.2	2.3	0.9
5	0.03	0.10	20.1	1.7	0.9
6	0.09	0.14	20.2	2.4	0.8
7	0.11	0.11	20.3	2.4	1.0
8	0.07	0.06	17.7	1.8	0.8

Table 2: Scores for the bridge images of Figure 7

Figure 8: Face image interpolated by Methods 1 to 8, reading from left to right. The correct viewing distance is 55 cm.

Method	Visual Difference Score		Signal-to-error	Human Ranking	
	Mode	Median		Mean	σ
1	0.37	0.48	23.6	3.6	1.1
2	0.38	0.41	23.9	3.6	1.1
3	0.35	0.30	24.1	4.6	1.1
4	0.02	0.06	25.5	2.8	0.7
5	0.02	0.04	26.7	1.1	0.4
6	0.13	0.10	26.8	2.5	0.9
7	0.03	0.09	27.0	2.5	1.1
8	0.02	0.03	23.4	1.9	0.6

Table 3: Scores for the face images of Figure 8

4 Discussion

For the bridge and face image, Methods 5 (B-spline interpolation) and 8 (edge-enhanced zooming) were preferred by the human panel, and had low visual difference scores. Furthermore the panels' choices would not have been predicted by signal-to-error ratio. For the text image the visual difference scores did not predict the panels preference. This may be because human observers are judging *readability* not image quality [1]. We also found a very simple method based on truncating the DCT coefficients [7] was quite effective, so for applications where computational complexity is important this may be a good choice.

It is of interest to formally compare the visual difference scores to the human panel. We do this via the robust Spearman rank correlation coefficient, r_s, [14]. Each observer produced a separate ranking so we have computed r_s between each observer and each of the error measures. Table 4 gives the mean of the rank correlation coefficient over all 13 observers.

	Mean Rank Correlation Scores		
	Visual Difference Score		
Image	Mode	Median	Signal-to-error
Text	0.58	0.66	0.85
Bridge	0.61	0.52	0.31$*$
Face	0.67	0.71	0.28$*$

Table 4: Mean Spearman rank correlation coefficients for each image and error metric.

The rank correlation coefficients have a zero significance statistic that is distributed according to a Student's t distribution with 6 degrees of freedom. We find that a correlation of 0.28 is roughly at the 50% significance level. What this means is that all the correlations in Table 4, apart from those marked with a $*$, are significantly different from zero.

The visual difference score appears to predict human performance for a variety of grey scale images (in this paper we show only two of the images we have tried), but does not work well for images where the observers may use additional interpretation to assess image quality. The text image presented here is an example of this, but we have also found some face images have produced a disparity between the visual difference score and human scores.

A further problem when testing resampling methods is the provenance of images. All of the images shown here were generated using known methods, but we have noted inconsistent results when using well known test images. We suspect that several of these images have been previously resampled.

Currently we are repeating the human tests using, what we think is, an improved method in which observers are presented with image pairs for a short duration and are asked to select the better of the two. A ranking is produced by sorting these pairwise comparisons.

[1] Given this observation it is surprising that papers on image resampling often use text images [12].

References

[1] J. Andrew Bangham, Richard Harvey, Paul D. Ling, and Richard V. Aldridge. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5(3):283–299, July 1996.

[2] J. Andrew Bangham, Paul D. Ling, and Robert Young. Multiscale recursice medians, scale-space and transforms with applications to image processing. *IEEE Transactions on Image Processing*, 5(6):1043–1048, june 1996.

[3] D.Fraser. Interpolation by the FFT revisited – an experimental investigation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(5):665–675, May 1989.

[4] Bernd Girod. What's wrong with mean-squared error? In Andrew B. Watson, editor, *Digital Images and Human Vision*, pages 207–220. MIT Press, 1993.

[5] G.Legge and J. Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70(12):1458–1471, 1980.

[6] H.Marmolin. Subjective mse measures. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-16(3):486–489, May/June 1986.

[7] I.Agbinya. Two dimensional interpolation of real sequences using the DCT. *Electronics Letters*, 29(2):204–205, January 1993.

[8] J.Daugman. Uncetainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensionals visual cortical filters. *Journal of the Optical Society of America*, A2(7):1160–1169, July 1985.

[9] J.Mannos and D.Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions on Information Theory*, IT-20(4):525–536, 1974.

[10] M.Unser, A.Aldroubi, and M.Eden. Fast b-spline transforms for continuous image representation and interpolation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(3):277–285, March 1991.

[11] M.Unser, A.Aldroubi, and M.Eden. The l2 polynomial spline pyramid. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 15(4):364–379, April 1993.

[12] M.Unser, A.Aldroubi, and M.Eden. Enlargement or reduction of digital images with minimum loss of information. *IEEE Transactions on Image Processing*, 4(3):247–258, March 1995.

[13] Allen B. Poirson and Brian A. Wandell. Appearance of colored patterns: pattern-color separability. *Journal of the Optical Society of America*, 10(12):2458–2470, December 1993.

[14] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C*, chapter 14, pages 640–642. Cambridge University press, second edition, 1996.

[15] P.Teo and D.Heeger. Perceptual image distortion. In *Proceedings International conference on Acoustics Speech and Signal Processing*, pages 982–986, 1994.

[16] S. Thurnhofer and S. Mitra. Edge-enhanced image zooming. *Optical Engineering*, 35(7):1862–1870, July 1996.

[17] C. van den Branden Lambrecht. A working spatio-temporal model of the human visual system for image restoration and quality assessment applications. In *Proceedings International conference on Acoustics Speech and Signal Processing*, 1996.

[18] Christian J. van den Branden Lambrecht and Joyce E. Farrel. Perceptual quality metric for digitally coded color images. In *Proceedings European conference on signal processing*, 1996.