Invited Paper

Limiting human perception for image sequences

Anthony Maeder, [(1)] Joachim Diederich [(2)] and Ernst Niebur [(3)]

(1) School of Electrical and Electronic Systems Engineering
Queensland University of Technology, Brisbane, Australia

(2) School of Computing Science
Queensland University of Technology, Brisbane, Australia

(3) Krieger Mind/Brain Institute
Johns Hopkins University, Baltimore MD, USA

ABSTRACT

Early vision processes, based on Human Visual System (HVS) performance, provide insufficient information for modeling our assimilation of image sequences (e.g. video). The use of a visual attention paradigm for modeling viewer response over time is advanced here. An "importance map" of the scene can be constructed using both spatial and temporal information. The image quality of an individual frame can be degraded significantly using the importance map to predict typical foci of attention. Knowledge of the whole scene can be built up over many frames, by accumulating details represented at low quality in areas identified by the importance map as warranting less visual attention. We conjecture some limitations on the image quality and provide synthesized examples of scenes coded using this model.

Keywords: human visual system, perception, attention, salience, video

1. INTRODUCTION

Considerable effort is often expended in producing high-quality visual data scenes to be displayed for appraisal by human viewers. Some examples are computer graphics generated scenes for visualization, and archived digital images of natural scenes. This effort is expended because it is widely assumed that fine detail in the visual data must approach as close as possible to reality in order to satisfy the viewer. This assumption may be inappropriate in many cases, for the two reasons discussed below.

First, the *purpose* for which the visual data is used may make the presence of arbitrary fine detail unnecessary for satisfying the viewer. For example, a 3-D volume rendering for scientific data visualization purposes can allow as much understanding of the data to be gained by the viewer if it is undertaken with fewer colours, simpler illumination modeling etc. The viewer is still able to extract the appropriate information from the poorer quality scene because the task involves abstraction. This prior expectation that the visual data is to be used for further abstraction can occur in many

0-8194-2031-X/96/$6.00

other situations: e.g. computer games, spatial databases, multimedia retrieval. In the case of visualization, the abstraction process has been recognised widely enough for a loose formalism to be constructed around it.[1] In the other cases, implicit recognition of abstraction occurs in the way the visual data is handled, e.g. map generalisation according to scale in spatial databases; progressive image reconstruction according to channel and platform characteristics in multimedia retrieval.

Second, the detail provided in the visual data may exceed the capacity of the means of *delivery* to the viewer. Such capacity limits may be imposed by the display device characteristics (e.g. pixel spacing, intensity and colour ranges), or by the limits of human perception inherent in the viewer. Display device characteristics tend not to be used as a prior basis for explicitly limiting image quality, as this would reduce platform independence of the application. Nevertheless, some upper limit on image quality might be adopted if a highest capacity display device can be assumed, such as HDTV or laser printer. In such cases, only lower quality images would be needed for other allowed display devices, and these images could be obtained simply by decimation or resampling, or by a more complex data structuring approach, such as multiresolution or progressive image coding.

On the other hand, human perception limits are universal in nature and are highly restrictive in some cases. Considerable savings in computational effort and in the volume of data stored or transmitted could be achieved if any detail beyond the perception threshold could be suppressed. The way in which image contents are assimilated is dependent on many factors, but is strongly influenced by the scene content which provides a visual context. This context sensitivity suggests that it is unnecessary to provide uniform visual quality over the whole image, as this would include much detail that was perceptually unimportant. By adaptively varying the amount of detail provided locally in a scene, in accordance with perceptual limits at that location, an image of uniform assimilated quality could be produced. Our work aims to provide a systematic framework in which such perceptual limits can be assessed and expressed.

2. VISUAL PERCEPTION

Aspects of HVS performance have been widely studied[2,3,4] and our knowledge of them is based on physical and experimental evidence from physiology, psychology and other branches of cognitive science. Many processes in early vision are reasonably well understood, as they are associated closely with the physical functioning of the eye and preliminary analysis of the signals provided by it in the brain. Later vision processes requiring more sophisticated brain activity for linking these signals together and incorporating other mental facilities such as memory, are less well understood. Nevertheless, it is regarded as fundamental that early vision processes provide the basic raw data upon which later vision processes operate, so a knowledge of early vision is crucial as the first step in understanding human visual perception. We shall consider briefly several well known aspects of early vision which impose limits on perception.

The physical structure of the eye is well known: light is focussed by the cornea and lens onto the retina, where approximately 100 million small, densely packed receptors sample the input intensity (rods) and colour (cones). This information is fed to ganglion cells after having undergone a significant redundancy reduction, as the optic nerve which links the eye with the brain has only about 1 million channels to convey details of the sensation at the eye. Redundancy removal is accomplished by masking and adaptation processes. Masking allows the presence of high contrast patterns such as edges and simple texture to be represented. The detection of these patterns is associated with the contrast sensitivity function (CSF) characteristic, which determines how strongly edge or texture intensity transitions are detected. Adaptation ensures that the response of the receptors can cope with different levels of input values, e.g. widely varying intensity levels, but has the side effect that less detail can be detected for some inputs. For instance, colour sensation is lost at low intensities as the cones do not function at these intensities, and rods do not have appropriate differential spectral sensitivities. Both masking and adaptation are crude simplifications which can easily result in wrong conclusions if they are the only source of visual information, e.g. the well-known effects of simultaneous contrast and Mach banding inhibition.

For these and other early vision processes, it is possible to impose limits determined by physiology upon the image data, as a succession of independent steps.[5] For example, adaptation allows the intensity receptors to function over a dynamic range of about 10^3, so the range of intensities and colours used in the image may be correspondingly restricted. Maximum contrast sensitivity occurs between 1 and 5 cycles per degree of visual angle, so contrast and edge strength may be selectively degraded beyond this range. However, this approach does not take into account the context sensitive nature of later vision processes, which try to make sense of local features or regions of information in the scene. We have some simple models for this, such as the well-known "Gestalt" categories of visual organisation: proximity, similarity, continuation, closure and symmetry. These models do not provide sufficient information to allow precise analysis of how a particular image is assimilated; instead they identify significant components of the scene that may be perceived as separate entities. In this way they can be considered to enhance information provided from the early vision processes, to aid in determining the fundamental parts of the scene.

3. TEMPORAL PROCESSES

The visual perception processes identified above have all been described with reference to a still image. Rather than adding more independent processes to the set, motion introduces so much more visual information that it affects many other aspects of visual perception very strongly. The previous model of perception limits obtained by simply combining individual perception processes is no longer valid when motion is included. We need to consider the particular effects of motion, both of the eye relative to the image and of the image contents internally.

Motion of the eye arises because the resolution of the retina varies: it is highest across the centre few degrees (fovea) and drops off rapidly towards the periphery. In order to acquire detailed visual information over the whole scene, the eye is repositioned by small saccadic movements 2-5 times every second. A saccade moves the eye to foveate on a place in the scene where significant detail occurs. The HVS then devotes most effort to processing of the data from this area, corresponding to the overt focus of visual attention, until the next saccade. A consequence of this saccadic acquisition of information is that the detail perceived by a viewer, in a scene presented for only a short period, will be concentrated in the immediate surroundings of those points which have served as foci of attention. If these points could be predicted beforehand, or detected when the eye moves, high quality image detail would need to be provided only there. However, no reliable model exists for predicting focus of attention reliably, and so this is an area of much ongoing research.

Motion of image contents occurs when features or regions of the image change, through their own movement or through movement of the observer (or camera). Rapid motion may result in a loss of detail due to blur of texture or edges. Common instances of image motion comprise combinations of rigid movements such as zoom, pan, rotate. More complex motion may involve non-rigid plastic deformations or catastrophic deformations. The eye is able to follow simple motion of almost 10 degrees per second in smooth pursuit, and so can retain the capacity to observe detail for a moving focus of attention. [6] Indeed, image motion tends to attract viewer attention and so it is easier to predict where the focus of attention is most likely to be directed by tracking the motion. [7] In highly structured sequences where the type and location of motion can also be predicted, a fixed focus of attention model can be adopted successfully (e.g. centred on the face for videophone images [8]).

Many experiments have been conducted on eye movement patterns since the pioneering work of Yarbus. Eye tracking devices can be used to sample the instantaneous position of the pupil (or retina). From this information, characteristic saccadic behaviour can be observed for different viewers and different scenes. It is conceivable that eye position could be measured rapidly enough to allow dynamic local image quality adjustments, so that higher quality was always present at the focus of attention. However, this may lead to distraction of the viewer due to effects of the changing quality on covert attention. It is also feasible to collect information on a viewer's eye position behaviour for given scenes. A system might then be trained to vary image quality according to predicted eye movements, based on these observations. Instances of the latter approach, using neural networks to undertake training, have proved successful in other computational situations involving user interaction. [9,10] These could be extended to construct hypotheses on differences between users, and so add further sophistication to the training procedure. The approach described below does not implement dynamic quality variation, but could be extended in principle for this purpose.

4. IMPORTANCE MAPS

Predicting the focus of attention either statically or dynamically requires a particular position within the image to be specified. Positions of greatest potential interest can be computed by various means: we have adopted an approach based on assessing every point in the image relative to a chosen factor, resulting in a pseudo-image or "map" which provides an interest value at every (x,y) position. By combining such interest maps for different fixed factors corresponding to various perception processes, an overall "saliency map" can be formed. [11] A neural network based version of this approach has recently been demonstrated successfully. [12] Our simplified version considered here avoids feedback by computing linear combinations of the factors, once they have been rendered mutually conformant. To distinguish this version from the neural based model, we term our resulting map an "importance map". This simplified approach has been effective in modeling structural information in colour images [13] and variable lossy compression criteria for natural scenes. [14]

An importance map is constructed by first selecting visual perception factors, which approximate the information extracted from the image by both early vision and later vision processes. Some common factors are edge strength, texture energy, contrast, colour variation, homogeneity, global probability. Values for these factors are computed over a neighbourhood in the case of local factors, and over a region or the entire image in the case of global factors. The factor values need not be computed by linear formulas, e.g. the maximum or minimum of a local property may be used. Each factor is individually normalised to the range [0,1] and linearised via a process akin to histogram equalisation. The final value for each position in the importance map is computed as the average of each factor value in that position. This overall procedure is summarised in Figure 1. Importance values may be averaged further to provide neighbourhood importance rather than point importance. The importance values may be further quantised to establish importance classes, e.g. percentiles of the importance distribution may be established, and the importance values set to represent one of these percentiles.

5. MAPPING SEQUENCES

The importance map concept has been generalised to cater for image sequences. This was accomplished by using a static importance map of the first image in the sequence to predict the initial points of importance, and then modifying this map over time according to image motion information. A motion map was combined with the static importance map to increase the importance of high importance points undergoing motion, and to decrease the importance of high importance points not undergoing motion or low importance points undergoing motion. The motion information is monitored to ensure that consistent smooth motion occurs at high importance points. When a major motion change is encountered (e.g. new sequence or catastrophic change), the static importance map for that image is used.

Point or neighbourhood importance values can be used to determine local image quality, according to limits assumed for assimilation at points of highest and lowest importance. We estimate the range of these limits as follows. Assuming n shifts of attention per second and a sequence of k frames, displayed at v frames per second, we expect $s = (n.\, k)/v$ shifts in total, and so we might choose at most s classes of importance values. In the first image in the sequence we represent neighbourhoods belonging to the most important class with greatest visual quality, and those of less important classes with correspondingly less quality. Over time the quality of these lesser classes is increased slowly and fairly uniformly, as bandwidth allows.

Figure 2 shows examples of frames from two image sequence, where 8x8 importance neighbourhoods were used. In both cases, 4 shifts of attention were allowed, so the importance values were divided into 4 classes at the quartiles of the distribution. The resulting importance maps are shown in Figure 3. The images were encoded using JPEG, with the quartile quality as measured by PSNR reduced by 1dB, 2dB and 4dB for each successive class below the highest, which was set at 22dB in both cases. In the landscape scene, small scale motion in the water would lead to slowly increasing importance values there. In the face scene, motion of the eyes and mouth would lead to increased importance values there, alternating between whichever was currently undergoing the greatest motion. Both images are represented at more than twice the compression rate of the full resolution versions (i.e. half the number of bits per pixel), which would have been used under conventional circumstances where a uniform quality at the highest quality level required would have been adopted. The uniform highest quality versions are shown for comparison.

6. CONCLUSION

A basic method for estimating the importance of points in an image sequence in contributing to viewer understanding has been presented. Using this method, portions of the images can be selectively constructed at high or low visual quality, in harmony with the characteristics of the HVS. The application of the method for two different examples shows plausible variable quality images which do not appear significantly degraded when viewed while changing over time. The method is currently being considered for use in situations where feedback from users is available, in conjunction with eye-tracking measurements, so that further refinement of the selection of attention positions and levels of quality can be addressed. This work was supported in part by the Australian Research Council.

7. REFERENCES

1. P.R. Keller & M.M. Keller, *Visual cues: practical data visualization*, IEEE Press, Los Alamitos 1993.

2. K.T. Spoehr & S.W. Lehmkuhle, *Visual information processing*, W.H. Freeman, San Francisco, 1982.

3. R. Sekuler & R. Blake, *Perception*, Alfred A. Knopf, New York, 1985.

4. R.L. DeValois & K.K. DeValois, *Spatial vision*, Oxford University Press, New York, 1988.

5. W.E. Glenn, "Digital image compression based on visual perception", in, A.B. Watson, (ed), *Digital Images and Human Vision*, pp. 63-71, MIT Press, Cambridge MA, 1993.

6. M.P. Eckert & G. Buchsbaum, "The significance of eye movements and image acceleration for coding television sequences", in: A.B. Watson (ed), *Digital Images and Human Vision*, pp. 89-98, MIT Press, Cambridge MA, 1993.

7. L.B. Stelmach, W.J. Tam & P.J. Hearty, "Static and dynamic spatial resolution in image coding: an investigation of eye movements", *Procs. SPIE*, 1453, pp. 147-152, 1991.

8. D.M. Bell & A.J. Maeder, "A progressive human face image archiving and retrieval system", *Procs. SPIE*, 2606, pp. 101-110, 1995.

9. J. Diederich, A. Thummel & E. Bartels, "Recurrent and feedforward networks for human-computer interaction", in: B. Neumann (ed), *ECAI-92 10th European Conference on Artificial Intelligence*, pp. 206-207, John Wiley & Sons, Chichester, 1992.

10. J. Diederich & M. Wasserschaff, "Recurrent neural networks for sequence production", *IJCAI-93 International Joint Conference on Artificial Intelligence*, Morgan Kaufman, Los Altos, 1993.

11. C. Koch & S. Ullman, "Shifts in selective visual attention: towards the underlying visual circuitry", *Human Neurobiology*, 4, pp. 219-227, 1985.

12. E. Niebur & C. Koch, "A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons", *Journal of Computational Neuroscience*, 1, pp. 141-158, 1994.

13. A.J. Maeder & B. Pham, "A colour importance measure for colour image analysis", *1st IS&T Color Imaging Conference: Transforms and Transportability of Color*, pp. 232-237, Scottsdale AZ, 1993.

14. A.J. Maeder, "Importance maps for adaptive information reduction in visual scenes", *Procs. ANZIIS-95 3rd Australian and New Zealand Conference on Intelligent Information Systems*, pp. 24-29, Perth, Australia, 1995.

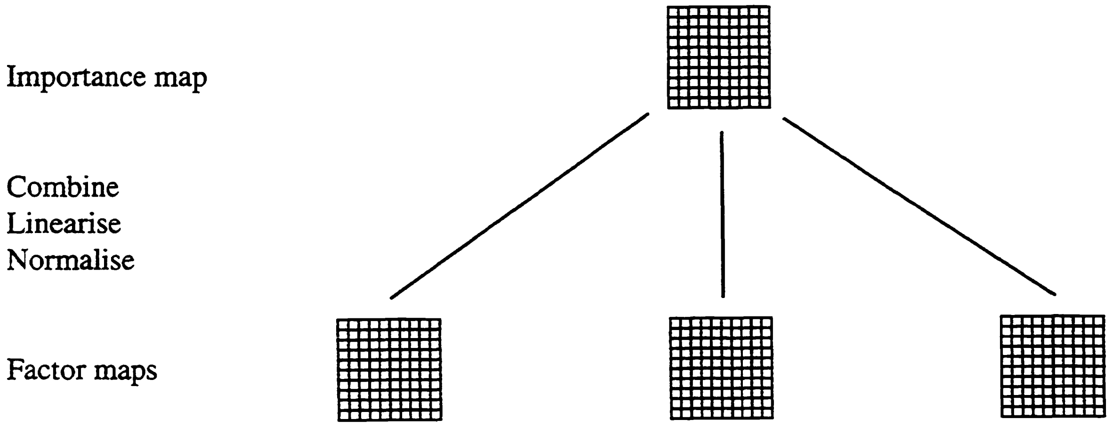

Figure 1: Overall importance map construction procedure.

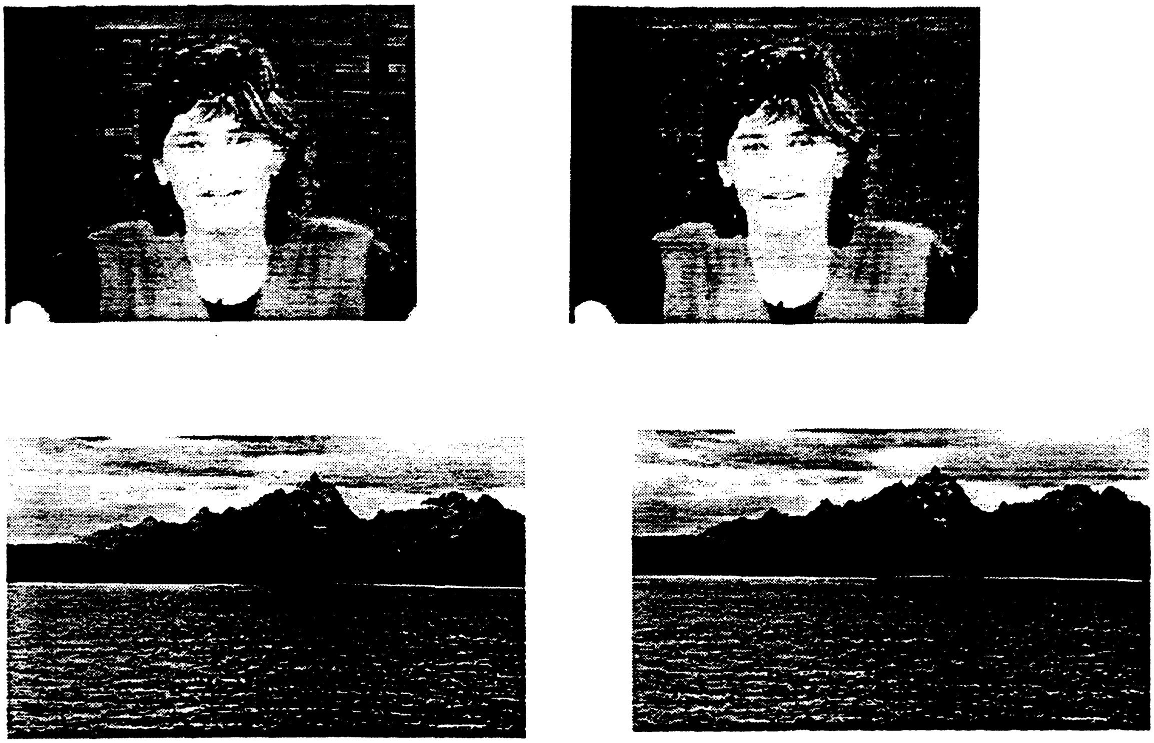

Figure 2: Frames from natural scene and videophone sequences represented with importance weighted visual quality (left) and uniform highest visual quality (right).

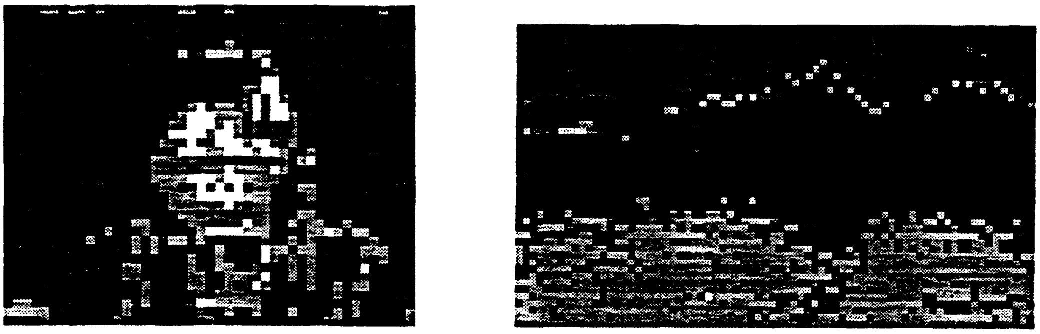

Figure 3: Importance maps used to construct frames in Figure 2.