SOME WIDELY HELD MISCONCEPTIONS ABOUT HUMAN VISION AND IMPLICATIONS ON PRACTICAL HIGH FIDELITY IMAGE PROCESSING.

I.Overington* and M.S.Overington**

*** Independent Consultant (Retired after 40 years at BAe), United Kingdom**
**** Independent Consultant, United Kingdom**

INTRODUCTION.

It would appear from certain recent general reference literature publicised by the IEE - Russ (1) - that there are a number of popular misconceptions about the performance capabilities of human vision which are prevalent in the image processing community. Possibly partially as a result of these misconceptions, there are also apparently a number of gross misconceptions about what is readily possibly in terms of simple, yet high fidelity, image processing. In this paper we would like to address two of what we feel to be the most important of these misconceptions.

Firstly, it seems to be widely considered that "human vision is primarily qualitative and comparative" (1). This appears to lead on to an assumption that one cannot learn a great deal about really high fidelity objective image processing by studying the mechanisms of human vision. It seems to us that it should not take a great deal of careful thought to realise that such a concept cannot be totally true. If it were, how then are we readily able - usually almost instantaneously - to recognise such things as

- subtle small shapes (including alpha-numerics etc.),
- strictly local straightness or degree of curvature of lines & edges,
- smoothness or roughness of surfaces

etc., all to a surprisingly high level of fidelity relative to the sampling intervals set by the retinal receptor spacing - see Overington (2) for a collection of human visual threshold limits? By and large, many of these tasks require local image interpretation capabilities which are considerably finer than the foveal retinal receptor spacing (where the utmost criticality of vision has to be carried out). Granted, such capabilities have to be ***learnt*** in the first instance, but we would argue that such learning is much more a limitation of the ***interpretative*** parts of the brain than the actual hardwired visual image processing networks.

Secondly, it appears to be generally accepted that stereo fusion has to be a laborious & difficult task (at least, that is generally ***assumed*** to be the case from an image processing standpoint - e.g. (1), Wei Sun and Sweeting (3)). But stereo fusion in human vision, together with relative local depth estimation, is being carried out by our visual systems virtually all our waking lives, both being accomplished effectively instantly and with very high fidelity! We therefore argue that the assumption that such stereo tasks must be difficult cannot be wholly correct.

It will be shown that, if one can unravel some of the (rather subtle) image processing operations occurring within the human visual tract and simulate them, then, particularly with the aid of fast modern PC's, it becomes possible to carry out these assumed to be difficult or laborious image processing tasks very quickly & easily (even with very compact software & simple digital cameras).

In one short paper we cannot even ***begin*** to ***explain*** what we believe to be the necessary subtle processes employed in human vision or the details of implementation of suitable simulation of such processes. However, these have been documented at length in previous publications, particularly Overington (4) (which it seems may not have been adequately publicised). Rather, in the present paper we must limit ourselves to a few summary statements, followed by a few practical demonstrations of how adequate simulations of human vision can provide very powerful results (which also have been found largely to match the limitations of performance found for human vision e.g. (2 & 4)).

A SUMMARY OF SOME IMPORTANT IMAGE PROCESSING FACETS OF HUMAN VISION.

The following, in summary, are a short list of what we believe to be the more important facets of image processing by human vision which enable the extremely high fidelity performance to be achieved thereby.

- The incoming image is blurred quite strongly relative to the foveal receptor spacing (the image falling on the retina has a roughly bell-shaped point spread function with a standard deviation of something like 1.0 to 1.3 receptor spacings). This level of blur is one which would be considered totally unacceptable from an engineering standpoint!
- Although not directly relevant to the examples to be given in the present paper, the eye lens is totally uncorrected for colour, with the blue end of the spectrum being wildly out of focus
- The retinal receptor matrix is very roughly ***hexagonal***, as opposed to the generally accepted

modes of ***instrumental*** image processing being square / rectangular. This has important impact on both connectivity processing for edges / lines and ability to determine local edge / line orientation.

- By far the majority of the ongoing image processing beyond the retinal receptors is by local or more extended (pooled) difference processing - that is, almost all information being transmitted onward to the cortex is in terms of local ***and*** broader ***difference*** signals.
- The difference signals are sensed for onward transmission in ***two halves*** - essentially the positive and negative parts of local differences. This splitting of the difference signals into two halves is not only convenient from a ***practical*** standpoint (avoiding any need to retain floating zeros) but also has the subtle property of permitting later reconstruction of both local ***second*** difference data and approximations to local ***first*** difference data.
- The ongoing difference data are subsequently pooled by elliptical neural pools set at various orientations, such that outputs from that level of processing contain powerful data about local ***orientation*** of lines and edges.

The total result of all the foregoing is that, at a high cortical level, at any one locality there potentially exist local strength and orientation data from various areas of processing pools as centred on each & every retinal pixel. It has been found that the foregoing basic set of processes are capable of providing and commensurate with a wide range of experimentally determined visual performance limits / capabilities.

COMPUTER SIMULATION.

With the foregoing list of incremental visual processes, it might be thought that any attempt at total simulation of same would be doomed to failure. However, as discussed in very considerable depth in (4), this has proved not to be the case. To the contrary, it has proved possible to construct a composite suite of image processing algorithms which are capable of simulating the majority of low level visual tasks (that is, other than the actual ***interpretation*** in such cases as legibility tasks) in a software package taking up less than 1MB of computer space and processing rather large images from modern digital cameras (up to at least 2400 x 1800 pixels) in a matter of a few seconds at most. Under normal usage, this suite of processes

- senses significant local edges / lines (including local motion or stereo disparities for temporally or spatially separated image pairs),
- segments the image based on the sensed significant local edges / lines,
- generates data files containing the characteristics of both segmented regions and region boundaries (and the interactions between the two).

Furthermore, provided with the rich output data for both segmented regions and local region boundaries, it has also been shown capable of providing such things as local boundary curvature trends, mean disparity data for best fusion of pairs of images exhibiting substantial mismatch and complete stereo surface maps.

SOME PRACTICAL EXAMPLES.

There follow examples of

- high fidelity region segmentation & boundary derivation from coarsely pixelated image data,
- determination of the correction necessary for ‘best’ fusion of a pair of images exhibiting gross mismatch,
- generation of a complete two-dimensional surface depth map from a stereo pair of images of a locally distorted steel plate.

High fidelity region boundary derivation from coarsely pixelated inputs.

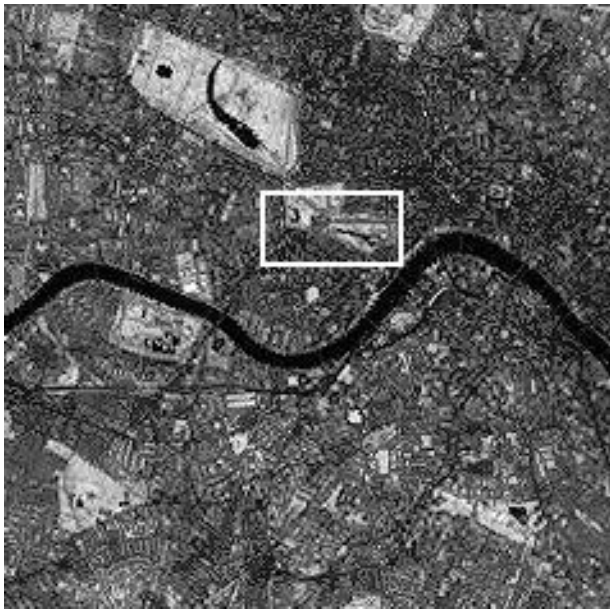

Fig. 1. Part of an IR satellite photograph showing Central London.

As part of a recent exploration of what it was possible to extract from typical satellite images, we were able to study a satellite image of the London area (figure 1). We were struck, on first looking at this image, how much detail it was possible to discern visually on the computer monitor screen. In particular, we noted that, in the area around Buckingham Palace (boxed), it was readily possible to discern the lakes in Buckingham Palace grounds and also in St. James Park. When viewed at 'Actual Size' and at normal reading distance (25cm) we also noted that the ***shapes*** of the two lakes as discerned were closely similar to those shown on a large scale map

of Central London. However, when we ***enlarged*** the appropriate portions of the image to obtain a clearer view (figure 2), we found that, rather than discerning ***more*** information, we actually discerned less! With hindsight, it was realised that here was a good illustration of ***one*** of the little appreciated capabilities of human vision - that of being able to 'in-fill' intelligence over and above that which is actually there. One could then pose the question "Is this in-filling a matter of guesswork on the part of our brains or is at least ***some*** of it genuine?".

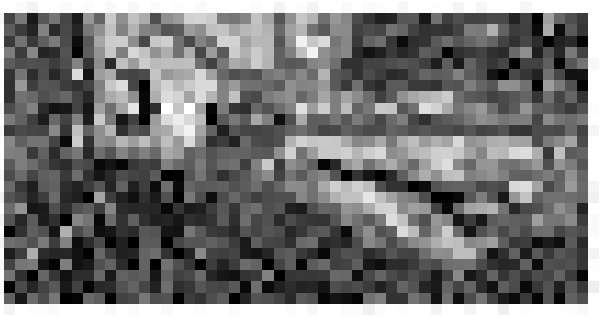

Fig. 2. Enlargement of the boxed area of figure 1.

It seemed that here was a particularly interesting problem to throw at our vision simulation, so an appropriate portion of the input image was selected and offered to the simulation software to see what was made of it. The result was as shown in figure 3, where it can be seen that, for both lakes, the said lakes are satisfactorily segmented as separate regions and the region boundaries show a strong similarity to the shapes shown on the map of London, whereas the original sampled fragments show little form at all!

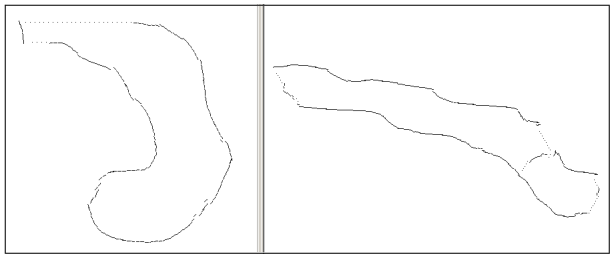

Fig. 3. Output profiles of the two lakes in figure 2.

Derivation of correction data for 'best' fusion of a pair of mismatched images.

One of the capabilities of human vision which has been simulated is the ability, on a pixel to pixel level, to derive the local magnitude of stereo disparity at significant brightness discontinuities to a small fraction of a pixel simulated is the ability, on a pixel to pixel level, to derive over a range of about +/-2 pixels disparity. Furthermore, in practice this can be thought of as being potentially available for ***any*** scale of image. Hence, if an original image pair of high resolution is scaled down greatly (by 'blocking'), the much lower resolution image pair will ***also*** be capable of generating local magnitudes of stereo

Fig. 4. A stereo pair of views of a typical rural scene.

disparity to a small fraction of ***its*** pixels. But this means that it is capable of providing at least an ***approximate*** estimate of the ***average*** mismatch over a much larger range of mismatches of the original image pair than possible at full resolution. By using this first approximation one can readily adjust the relative overlay of the original images roughly, after which one can equally readily reprocess using rather ***less*** scaling down, thereby generating a secondary correction for the original mismatch. Figure 4 is from an original pair of images of a typical rural scene, the images having a resolution of 2400 x 1800 pixels and with a horizontal mismatch of some 160 pixels. It has been found possible to determine the 'best' fusion correction for this image pair by such iterative processing in as little as ***three*** iterations (in this instance using 100:1, 50:1 & 24:1 scaling down - e.g. figure 5). Such processing can be carried out, even using manual settings and interventions, in as little as a few minutes. It is believed that, with fully automated software for this specific task, the actual processing time for a pair of 2400 x 1800 images should be at most a few seconds on a typical modern PC.

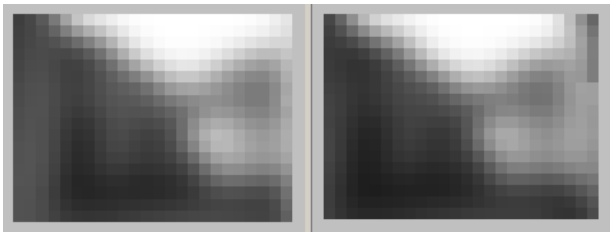

Fig. 5. Greyscale 100:1 reductions from figure 4.

3D depth mapping.

Early in 2002 we were presented with what seemed, by any yardstick, to be a very difficult stereo analysis task. In fact, although we claim that the human visual system is capable of quite critical stereo depth sensing, this ***particular*** task was even stretching that somewhat. A steel plate had been subjected to two local indentations which were of the order of 2 or 3mm deep (figure 6) and it was required to map the depressions in 3D to a depth accuracy (ideally) of 0.01mm. We knew that in theory we should be able to derive such a sensitivity of mapping (with everything going for us) if we viewed the plate from a distance of around 20cm using a stereo baseline of some 8cm. However, in practice various problems had to be overcome, even presupposing that our vision simulation

software could be shown to operate according to theory for stereo.

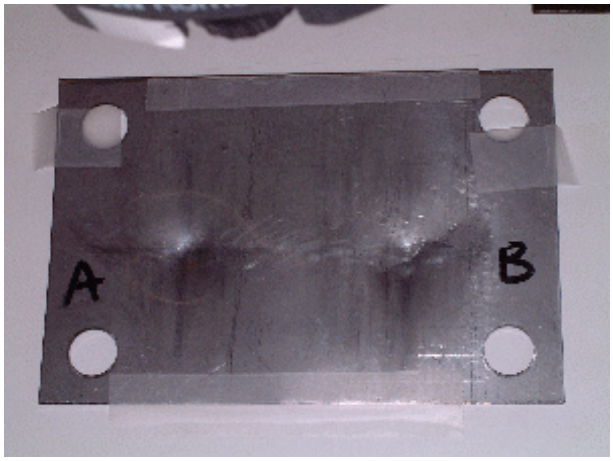

Fig. 6. View of steel plate with depressions.

It was found to be imperative that the alignment of the camera for recording from the two stereo viewpoints was highly accurate and it was also realised that, with such a large stereo baseline relative to the viewing distance, the toe-in necessary for the two views grossly contravened the normally accepted small angle approximations used for a wide range of optical considerations, resulting in significant & inverse trapezoidal distortion of the two images to be paired. Notwithstanding this, it was found possible to develop a breadboard rig on which to mount the camera in an adequately stable manner and it was further found possible to extend our software suite to produce appropriate (sub-pixel) adjustments to the two distorted images. Finally, since the plate had no edge structure of its own, we elected to generate an edge pattern by the simple technique of projecting a bar pattern onto the plate surface. Since this (fixed) pattern was to be viewed from the two different viewpoints, it would behave in exactly the same way as real edges in the object as far as the camera records were concerned.

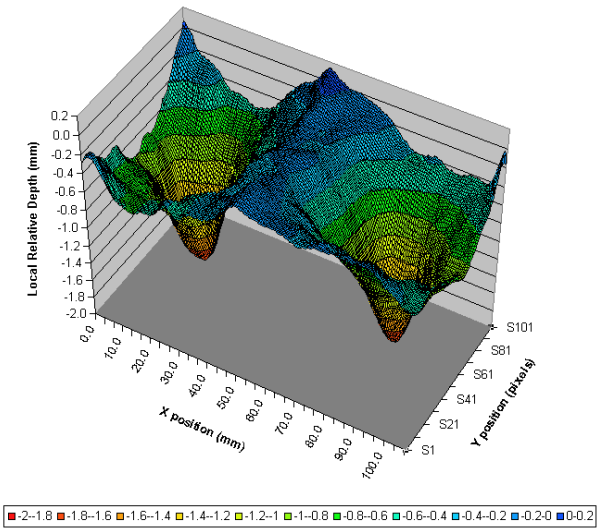

Fig. 7. 3D stereo depth map derived from the central portion of figure 6, *Y-scale: 1 pixel = 0.56mm*.

The resulting stereo pair of images were then offered to the vision simulation software and processed to extract, pixel by pixel, the local stereo disparity. This resulted in a rich tabulation of stereo disparity along each of the pattern edges. From this edge disparity pattern, a 2D continuous disparity map was developed, which was then converted into a true 2D distribution of local depth by means of the geometry of the viewing platform & camera. A 3D view of the resulting depth distribution is shown in figure 7.

CONCLUSIONS.

It is considered that the various practical examples of what can be accomplished readily with software based on the image processing techniques believed to be used by human vision should provide ample evidence of the gross misconceptions of capabilities of human vision and image processing based upon it.

A limited version of the simulation software has already been made commercially available for non-contact measurement of stress/strain relationships in fine textile yarns - Hearle *et al* (5), Overington & Overington (6).

We believe that there must be a considerable number of other potential applications for these forms of simple, yet high fidelity image processing and would be pleased to hear of any. It is perhaps worth stressing that such other potential applications in principle include 2D greyscale or pseudo-colour images created from such things as X ray radiography, ultrasound imagery and MRI scanning, amongst others.

REFERENCES.

1. Russ J., 2002, “The Image Processing Handbook - Fourth Edition”, CRC Press.

2. Overington I., 1976, “Vision and Acquisition”, Wiley.

3. Wei Sun and Sweeting M., 1999, “In-orbit results from UOSAT-12 earth observation minisatellite mission”, IAA-B3-0305P, Surrey Space Centre, University of Surrey, UK.

4. Overington I., 1992, “Computer Vision, a unified, biologically-inspired approach”, Elsevier B.V.

5. Hearle J.W.S., Overington I. and Overington M.S., 2001, “Non-contact strain and deformation measurement in yarns, ropes and fabrics”, IFAI.

6. Overington I. & M.S., 2002, Optical Extensometer Analyser (OEA), www.msoverington.co.uk.