A first demonstration of the capability of the Improc & Comvis suite, followed by coarse Steran whole frame mean stereo disparity analysis against real world scenes.

Introduction.

After recently becoming aware of a widespread documented belief that fusion of stereo pairs of images or other pairs of displaced images was a "laborious & difficult" task, it was considered time to attempt to demonstrate the ***simplicity*** of such tasks by simulating human vision as conceptually discussed in Chapter 10 of 'Computer Vision - a unified, biologically-inspired approach'.

In preparation for attempting such a task, in the autumn of 2002, whilst on holiday in the Lake District, I took the opportunity of capturing a selection of rural views which were purposely taken with predominent overlap of field of view, but at the same time with a substantial mismatch of the image pairs. One such image pair was as shown in figure 1. These images were captured with a Fuji 4900Zoom digital camera having a resolution of 2400 x 1800 pixels. It is apparent visually (by looking at the relationship between the two walls in the foreground against the edges of the frame) that there is a ***substantial*** mismatch existing. A rough check on the magnitude of this by checking the pixel values of the two walls in the foreground using 'Picture Publisher' showed that this mismatch amounted to something approaching 200 pixels. It was considered that such a substantial mismatch was more than sufficient to use as a serious test of the practical capabilities of the computer vision software suite 'Improc' & 'Comvis' to generate required image pair fusion correction data when used in conjunction with the supplementary area averaging software 'Steran'.

Fig. 1a. Original scene (lefthand viewpoint). 2400 x 1800 pixels.

Fig. 1b. Original scene (righthand viewpoint). 2400 x 1800 pixels.

Summary of progressive fusion theory.

The theory of basically how human vision achieves almost instant fusion of stereo image pairs from our two eyes whenever we change our gaze is, according to my developed understanding, that each of our eyes generates a concentric set of brightness difference images of the incoming scene at various scales. The two sets of difference images are then merged / compared at higher neural level, such that at each scale, over a range of disparities which increases with the level of scaling, a highly accurate (sub-pixel) local measure of disparity is obtained ***relative to that level of scaling***. For instance, the ***highest*** resolution scaling, which is essentially based on Laplacian-like differencing at ***single receptor*** (or pixel) level permits estimation of local disparity to better than 0.1 pixels over a disparity range of +/-2 pixels. Similarly, larger scalings permit estimation of local disparity to lower accuracies, but over larger ranges - e.g. if working with a scaling of 10:1, this would mean estimation of local disparities to better than 1 pixel over a disparity range of +/-20 original pixels.

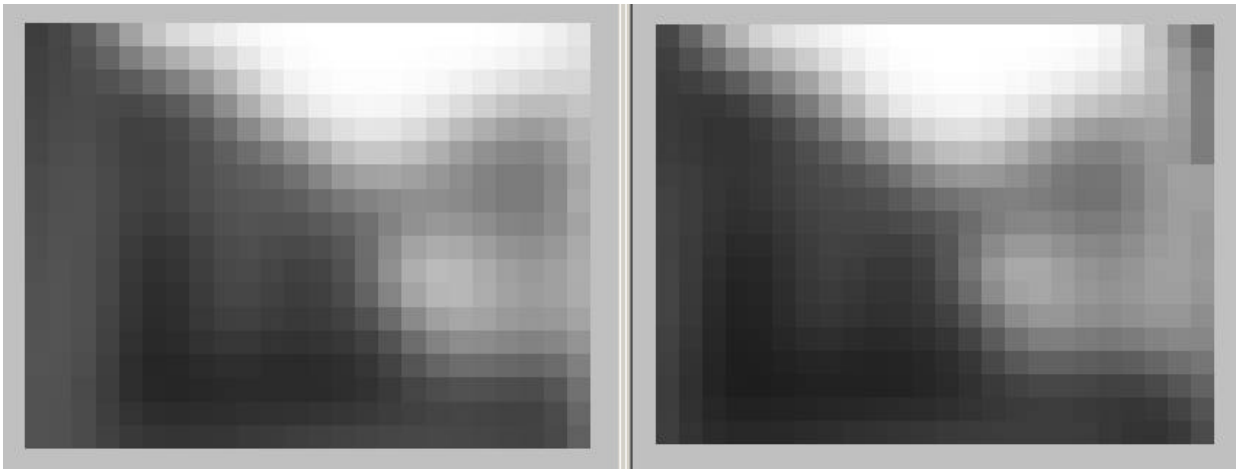
If the local disparities over a given wider part of the field of view are then pooled, and such a pooling is considered over the entire field of view, for any given scale a mean disparity can be very rapidly established ***for that scaling***. It is rather obvious that the large local disparities existing across a typical pair of images exhibiting gross mismatch cannot be handled successfully at the ***highest*** resolution, since in many situations this mismatch may well be several tens of pixels, together with possibly substantial ***differences*** in the mismatch for different parts of the field of view. However, for substantially coarser scalings, although the subtle ***details*** of the incoming images are gradually lost, there will nevertheless be local fluctuations in the 2D difference images which are intimately related to the original scene structure and, by definition, to each other. This is ***particularly*** true if, at each stage of scaling, the resulting image is effectively blurred.

In human vision, of course, ***part*** of the progressive scaling for the entire visual field is already provided at the retina by the progressive coarsening of the retinal receptor matrix away from the central, foveal region. It may well be, therefore, that for ***whole field*** disparity estimation and progressive fusion in human vision, this progressive, concentric coarsening provides an important part of the whole processing schema. However, although conventional digital imagery does not have such an input advantage, it is very easy to arrange to carry out progressive image scaling followed by sequential processing of individual scales of images, whilst making progressive adjustments as one moves from coarser to finer scales. Provided that one starts with an adequately coarse scale (such that the general levels of mismatch fall within +/-2 pixels at the coarsest scale), it proves to be more than sufficient to work with a series of scalings in ratios of around 2:1, making progressive adjustments to the input pairing as computed for each finer scale. Typically it should be sufficient to carry out no more than 3 or 4 iterations to arrive at a satisfactory overall fusion match.

Practical results.

As stated earlier, the image pair shown in figure 1 have a large horizontal mismatch, this being in the region of 150 to 200 pixels. Therefore, in order to provide an initial coarse scaling which should bring effective mismatch down to less than +/-2 pixels ***at the coarse scale***, it would appear to be necessary to carry out an initial processing on images which are scaled down by around 100:1 (greyscale is adequate). It is important that this scaling is carried out by ***blocking*** rather than ***coarse sampling***, in order that all subtle details are merged / averaged correctly in the resulting image. This scaled down image (of only 24 x 18 pixels!) should then be substantially blurred (to the same rules as have been found to be necessary to simulate high fidelity retinal imagery in human

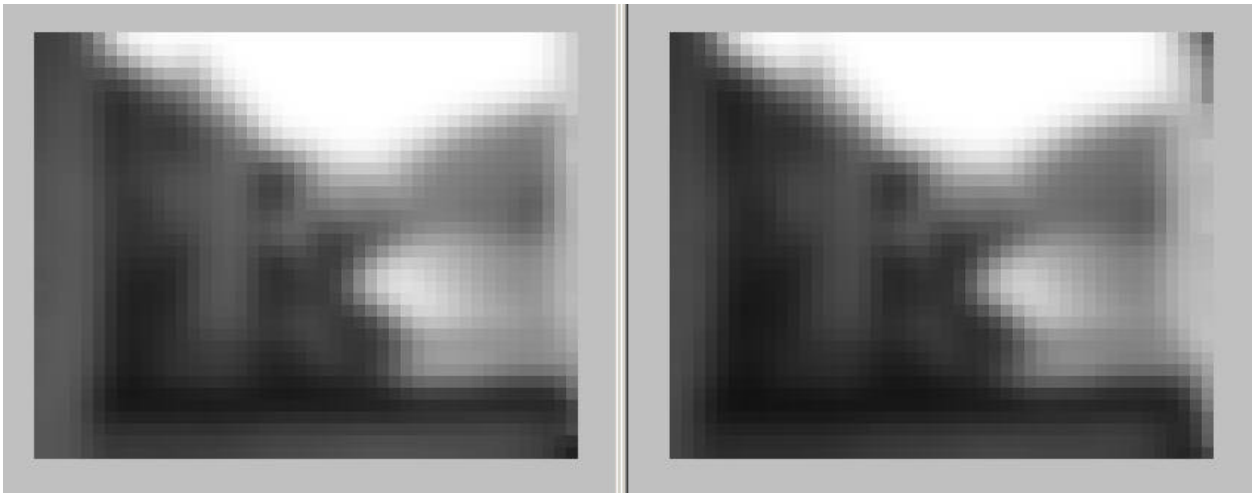
vision), resulting in a pair of scaled images as shown in figure 2. Here it can be seen that most of the detail to be seen in the ***original*** images has been lost, but nevertheless there ***is*** a recognisable scene structure existing which is comparable in the two images.



(a) lefthand viewpoint. **(b)** righthand viewpoint.

Fig. 2. 100:1 scaled greyscale images from figure 1 after applying standard blur.

On processing the pair of images in figure 2 through the computer vision simulation and then offering the output local disparity data to the supplementary disparity averaging process, a mean whole scene disparity figure of 1.01 was determined. This figure, multiplied by the scaling applied to arrive at the images in figure 2, should then be applied as a first order shift correction for carrying out a second paired frame processing (i.e. a shift of approximately 100 pixels on the lefthand image relative to the righthand image). Incorporating this 100 pixel relative shift, a second scaled image pair were generated, this time with a scaling of 50:1 followed by 'standard' blurring of the scaled images. This resulted in the image pair shown in figure 3, where it can be seen that a little of the original scene structure is being recovered, but the evident structure is still grossly distorted.



(a) lefthand viewpoint. **(b)** righthand viewpoint.

Fig. 3. 50:1 scaled greyscale images from figure 1 after standard blurring.

On processing the image pair of figure 3 through the computer simulation and then offering the output local disparity data to the supplementary software, a new mean scene ***residual*** disparity figure of 1.1278 was determined. This figure, multiplied by the secondary scaling of 50 (i.e. 56.35), provided a secondary shift component to be added to the 100 pixels already applied to the lefthand original image (i.e. a ***total*** shift of around 156 pixels).

Incorporating the total shift of 156 pixels for the original images, a third scaled pair were generated, this time using a scaling factor of 24:1 (24 rather than 25 for tidiness) followed by standard blur (figure 4). When ***this*** pair were processed as previously, a further residual disparity figure of 0.3377 was determined. This figure, multiplied by the latest scaling factor of 24 (i.e. approximately 8) needed to be added to the other two - i.e. 156 + 8 = 164.

(a) lefthand viewpoint. **(b)** righthand viewpoint.

Fig. 4. 24:1 scaled greyscale images from figure 1 after applying standard blur.

At this point it was deemed useful to explore rather more critically what the actual ***distribution*** of disparities in the original images was, bearing in mind that the scene comprised both comparatively close and comparatively distant image components. As a useful comparison of near and distant, three readily locatable edges were selected - one in the relative foreground (a house end) and two across the river in the middle distance (a second house end and a characteristic tree outline). It was found that the ***actual*** disparity for the near foreground was approximately 162 pixels, whilst the pair of disparities in the middle distance both measured approximately 168 pixels! So, even with only ***three*** (relatively coarse scale) processes it had been possible to estimate the shift necessary for approximate fusion to a mean value between near foreground and middle distance.

Since any further iterations could only generate further corrections of at most a few pixels - and since the variations between foreground and background were themselves several pixels - for ***whole scene*** best fusion it was considered unnecessary to proceed further. If, however, one wished to improve fusion for a ***particular part*** of the scene (at the expense of the rest), then it would be possible, using similar processing, to refine the (now essentially ***local***) fusion still further.

Conclusions,

From the foregoing, it is considered fair to conclude that recent published claims that determination of fusion for disparate pairs of images is a laborious and time consuming activity are grossly

incorrect. Even with ***manual*** interplay between stages of the processes described, the entire processing is possible in a matter of a few minutes. Of this, ***by far the majority*** of the time is taken up in setting and resetting the inputs for computation. It is believed that, with a small amount of effort, the processes can be automated, such that the entire fusion computation will be possible in a matter of ***seconds***!

I. Overington.
22nd January, 2003.