A SIMPLE METHOD FOR DEPTH RECOVERY FROM MULTI-CAMERA ARRAYS

G. L. Scott and S. Wossner

Cognitive Studies Programme
University of Sussex
Falmer
Brighton

ABSTRACT

This paper describes a simple, fast and robust method of obtaining a depth field from a large number of (closely spaced) views of a scene. The method - "synthetic aperture depth-from-focus" - involves superimposing images in such a way as to obtain a synthetic "blurred" image similar to that which would be obtained by a single camera with a very large aperture lens focussed for some particular depth. The full variety of depth-from-focus techniques is applicable to such a synthetic image or set of such images. In this paper however we explore a novel method which takes advantage of the anticipated symmetry of the "contrast profile" - the plot of contrast against the inverse depth of the effective focal plane.

I

A number of techniques - based both on single images and on a set of variously focussed images of the same scene - have been proposed for utilising "blur" to derive depth estimates. (e.g. [1], [2]). The difficulties encountered with such schemes in practice include:

1) The "parallax" afforded by the width of the lens is rather small - even when the lens stop is at full aperture - unless the scene being viewed is close. For out-door ranging with normal lenses the method is ineffective.

2) Multiple imaging (with a range of depth planes in focus) may be difficult in situations in which the scene or lighting is changing. D-from-f based on single images overcomes this problem but invokes strong assumptions about the nature of edges and is unrobust regardless of its theoretical viability.

Conventional *stereo* [3] overcomes these two problems but has some of its own!

1) The wider apart the cameras (to improve parallax) the more severe the correspondence problem.

2) Depths of edges oriented parallel to the displacement direction are not recovered.

3) Camera calibration, vibration and other "technicalities" can be major headaches in a variety of situations.

We have explored a system - aimed at eventual practical use - which seems to overcome the problems of both depth-ranging techniques. We use multiple cameras arrayed both horizontally and vertically which (notionally) image the scene simultaneously. Algorithms which combine the information from more than two cameras are not new [4],[5] but they tend to be tied to the stereo rather than the depth-from-focus "paradigm". We have experimented with two configurations: 9 camera positionson a 3x3 grid; and 25 camera positions on a 5x5 grid. Calibration errors of two pixels or so are detectable in our 3x3 imagery but the technique we use "loses them in the wash" so to speak. Depths of edge points in *all* orientations are recovered. Also, because of the small displacement which may obtain between any two *adjacent* cameras (compatible with maintaining high overall parallax) we are able to recover accurate depth on surfaces with fine texture.

A note: this is strictly an exercise in developing an engineering solution to the depth ranging problem. We do not suggest that Nature has erred in providing most animals with two or fewer eyes. It may be worthy of comment, though, that many animals - confronted with a non-trivial problem of ranging or camouflage breaking - make head movements which seem designed to provide images from a large number of viewpoints for "fusion".

II

To simplify analysis we consider only the case of a number of identical cameras mounted on a "base plane" with their optic axes orthogonal to the plane (i.e. parallel to each other). We take one camera to be at the origin (0,0) of the base plane and the image in this camera to be the "reference image". The disparity between the projection of a point in the reference image and its projection in a camera mounted at (X,Y) will be (-fX/Z,-fY/Z) where f is the focal length of the cameras and Z is the perpendicular distance of the point from the base plane. That is to say: if a point projects to image coordinates (x,y) in the reference image it will project to (x - fX/Z, y - fY/Z) in the other image. A point at infinity projects to the same coordinates in any camera mounted on the base plane so that if the image intensities from a number of cameras arrayed across the image plane are averaged these points will be "in focus". Points not at infinity will be "blurred" to a degree proportional to their inverse depth (1/Z) and proportional to the area over which the cameras are arrayed.

We can easily arrange for some depth other than infinity to be the "focal plane" by offsetting the images appropriately before averaging intensities.

AVC 1987 doi:10.5244/C.1.24

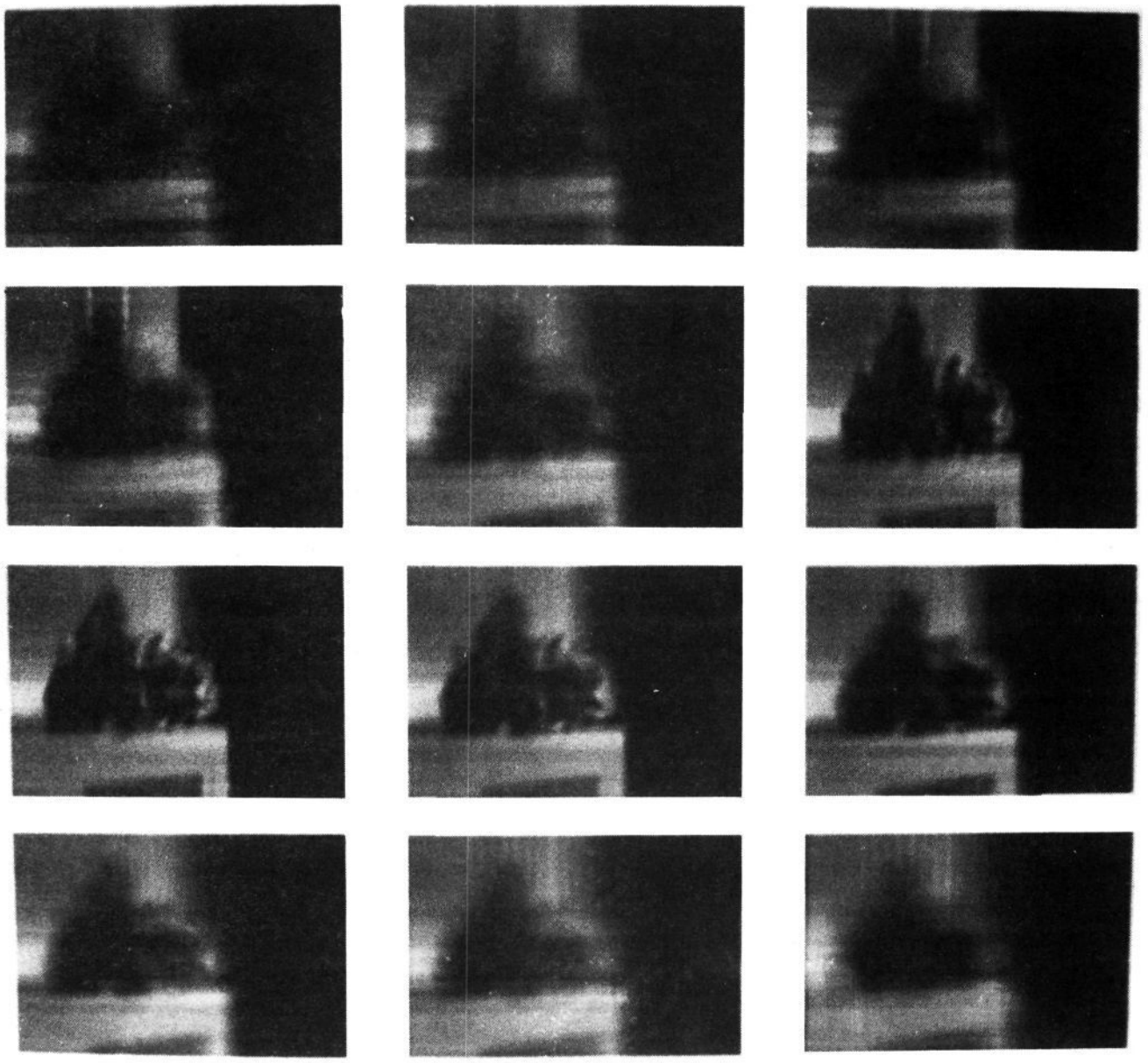

Figure 1.

Synthetic focus images for 12 depths obtained from 5x5 array.

An offset of (fX/Q, fY/Q) applied to the image of the camera at (X,Y) brings points at depth Q "into focus". A point of any other depth R will be "blurred" in the synthetic image by an amount proportional to $|1/Q - 1/R|$.

Figure 1 shows 12 synthetic pictures of a scene derived from a 5x5 array of images. The distance between adjacent camera positions is 2 cms. The original pictures were reduced to 256x256 for ease of handling. The step size is equivalent to one pixel for a camera adjacent to the central camera - two pixels for one of the outer cameras. The disparity range for objects in the scene is between 3 and 8 pixels. The offsets range from zero (top left-hand corner) to 11 (bottom right hand corner). Note the curious form of "aliasing" which occurs in some pictures due to the interaction between the two vertical gaps in the blind. Some attention needs to be given to camera configurations which are less regular than those we use - in order to defeat "autocorrelation" pheonomena such as these.

Simply offsetting images and adding intensities are procedures which may be undertaken very fast on some existing hardware (for example Datacube). We are keen to avoid procedures for depth estimation which tend to be time-consuming and we also wish to avoid procedures which can lead to serious errors *in practice* when conditions are not optimal. For these reasons we have eschewed both multiple cross-correlation and feature matching approaches. Under ideal conditions such methods would maximise extraction of depth information from the images. But we have rather a lot of images and are therefore in the luxurious position - relative to someone who owns a mere pair of images - of having information to burn.

III

The simplest contrast-measure that can be applied to an image patch is the *variance* of intensity over the patch - proportional to

$$\sum I^2 - N \sum I$$

where summation is over all pixels in the patch and N is the number of pixels. (We have experimented with other contrast measures - such as the remaining variance after first or second order surfaces have been fitted to the image function - but we do not report on these here.) If the patch is centered over a sharp edge or covers a uniform area of fine texture then we would expect contrast to be highest when the synthetic image is focussed for the depth appropriate to the patch. However, if the patch covers a uniform region with an edge just outside the patch then it will be seen that variance will actually to *minimal* when the image is in focus for the appropriate depth.

So we cannot obtain reliable depth estimates for the centre of a patch merely by recording the focal plane at which variance is highest. There are a number of possible approaches to depth-estimation based on contrast. The most obvious is to filter the reference image for "high interest" regions, viz. those which are indeed centred over an edge or filled with fine texture, and confine attention to these. This is liable to work well provided the reference image is sufficiently noise-free. The preliminary filtering is straightforward and need involve no more than convolution with a difference of Gaussians mask of a suitable scale. Near-zero responses indicate one of three things: centering over a well-defined edge, fine "random" texture, or uniform intensity. The last case can be neglected since high contrast measurements will not be observed in the image corresponding to any focal plane.

Note that it is actually easier to filter for "high interest" regions in the synthetic aperture situation than in the true depth-from-focus case. This is because we have a reference image which (relative to the degree of effective defocussing in the combined image) is "in focus" for all depths. The genuine depth-from-focusser has the tricky problem - when trying to locate genuine edges or textured patches - of knowing in which of his set of images to look for them.

IV

We do not pursue the "pre-filtering" approach in this paper but instead explore another possibility - one which does not involve reliance upon any of the individual images as a source of information about "high interest" patches. Some reflection will reveal that - for a region of uniform depth - the effects of "underfocussing" are identical to the effects of "overfocussing". This is to say that the "contrast profile" - the plot of variance against offset for any particular patch - should be symmetrical about the correct amount of offset. Also: the profile should ultimately decay in either direction as the patch becomes "hopelessly blurred" - though this will only be a completely reliable prediction where there is a sufficiently large number of cameras involved.

We are prompted to estimate the depth of a patch from the contrast profile by finding the phase and amplitude of the first Fourier component. The phase identifies the putative axis of symmetry of the profile and the amplitude provides a possible measure of confidence. In order for this method to work we must extend the blur profile sufficiently far in both directions for it to decay! This is certainly not the optimal method of finding the "most symmetrical" fit to any given profile - but it is fast and it works well over a range of interesting cases.

Figure 2 shows a variety of contrast profiles with the fitted sine wave of period equal to the width (= 12) of the range of offsets. These profiles were derived from the 5x5 imagery by estimating variance over 5x5 windows.

Some of the profiles do have the single peaked form characteristic of edges and fine texture (the very sharp peaks are "texture" peaks and the gentler sloped ones derive from edges). But some have the "twin peaked" character which indicates that they derive from regions close to, but not straddling, an edge. In these cases our technique goes some way towards uncovering the symmetry - whereas a simpler method such as choosing the highest peak would lead to serious errors.

V

For windows which give rise to single, clear cut symmetrical profiles our method yields rather accurate estimates of true depth. In the case of well-defined edges this accuracy could probably be

raised, with some sophistication, to the level attainable from well-calibrated conventional stereo based on matching edglets localised to sub-pixel accuracy. More interesting, however, is the performance on complex fine texture which is by and large better than the performance on edges. It is in such regions that conventional stereo suffers from its worst errors.

Where the contrast profile is symmetrical around a dip or a minor peak the disparity estimate yielded by our method may not be very accurate. But note: in the stereo case no estimate is yielded by matching in such cases - since there is not token to match. Depth in such cases is usually estimated by propagative methods. Such a "smoothing" or "neighbourhood support" process is in fact implicit in our technique. A point close to, but not actually on an edge will tend to receive the same disparity value as a point on the edge.

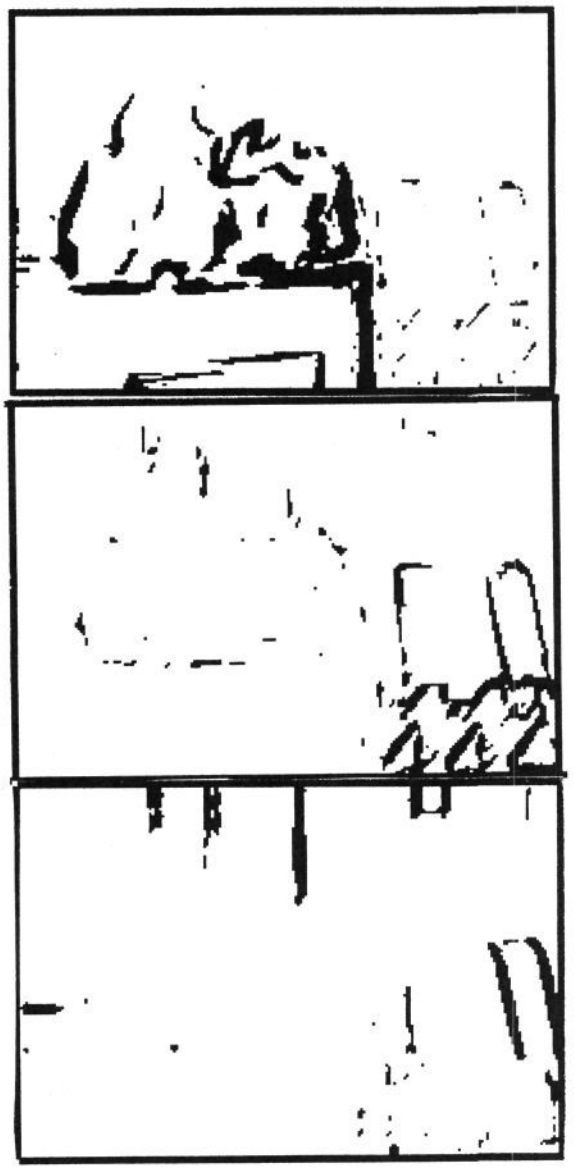

Figure 3: Recovered depths divided into three groups (5x5 imagery)

To illustrate the performance of our algorithm we have split the disparity space into three (figure 3). The central slice has a width of 1.5 pixels of disparity. Results shown have been thresholded (on the amplitude of the fitted sine curve). Most of the points pertaining to the crumpled paper and laser printer (for that is what they are!) fall, correctly, in the first "near" slice. The packet of Xerox paper (bottom right corner of the image) falls correctly in the second "middle distance" slice. Background features - corners, picture on the wall, window-sill, gaps in the blinds - fall appropriately into the "far" slice. The computer terminal (immediately behind the Xerox paper) has been split "middle distance" and "far" - a situation which conforms with its physical location.

A similar three-slice representation is shown in figure 4 for another scene. In this instance the array is 3x3, the images were reduce to approx. 128x128, and the variance was measured across a 3x3 window and the results have been cleaned with a "neighbourhood support" operation.

VI

The research reported here arose from an undergraduate project carried out by the second author and at present is not continuing. Questions concerning the number and configuration of cameras, fast computation and other matters need to be addressed before it will become clear whether "synthetic aperture depth-from-focus" constitutes a basis for robust, practical depth estimation.

REFERENCES

[1] Pentland, A.P.(1985), "A new sense for depth of field", *Intl Joint Conf on AI,* Morgan Kaufman

[2] Subbarao, M. (1987), "Direct recovery of the depth map", *Technical Report 87-02* Computer Vision and Graphics Laboratory, State University of New York, Stony Brook, N.Y.

[3] Pollard, S.B., Mayhew, J.E.W. and Frisby, J.P. (1985), "'PMF': A stereo algorithm using a disparity gradient limit." *Perception*

[4] Moravec, H.P. (1977), :Towards automatic visual obstacle avoidance", *IJCAI-77 no 5*

[5] Ayache, N. and Lustman, F. (1987), "Fast and reliable passive trinocular stereovision", *First Intl Conf on Computer Vision* p 422

Figure 3: Sample synthetic image from 3x3 array and separation into three depth "slices".

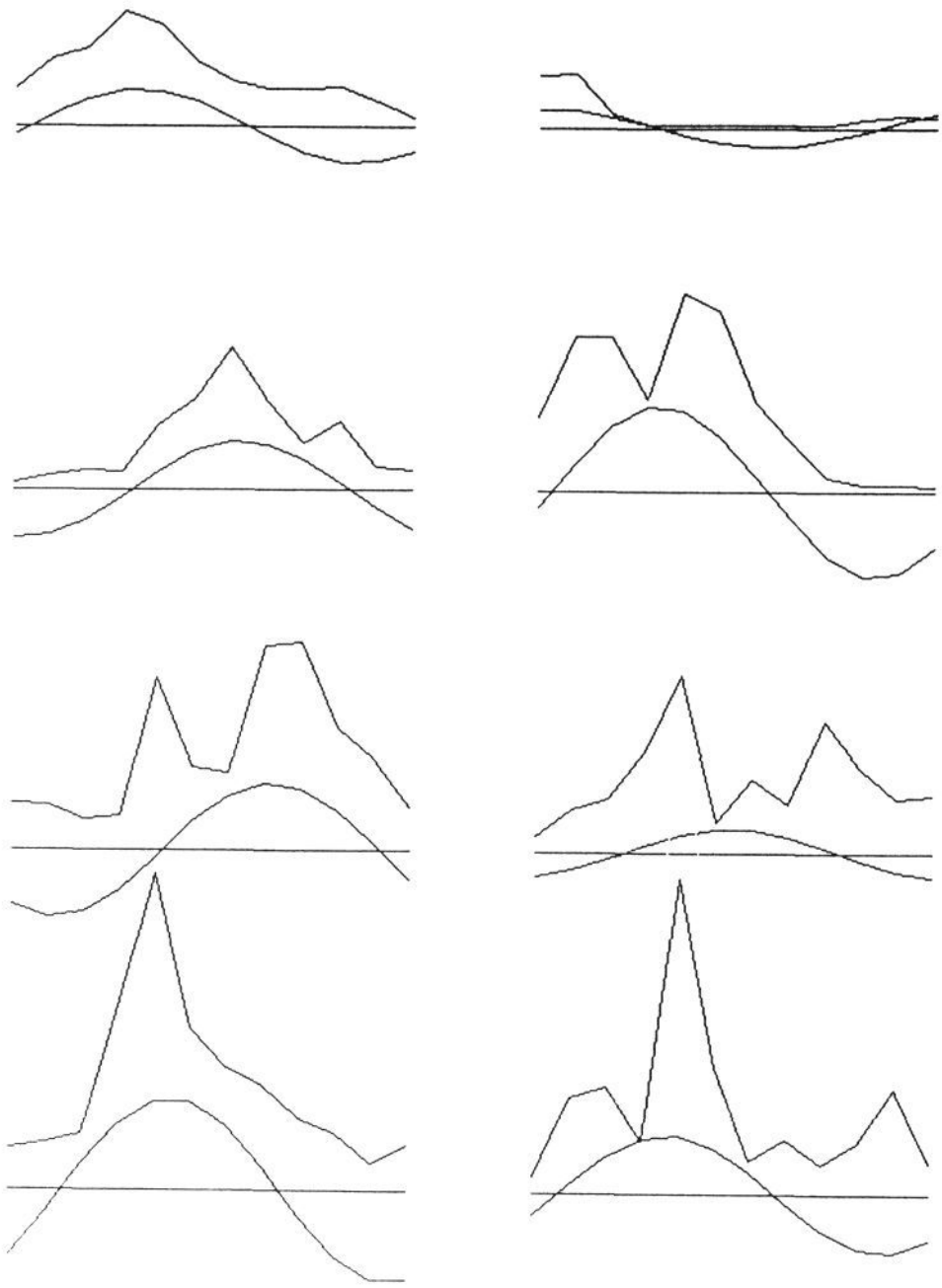

Figure 2: Contrast profiles and first Fourier component