Stereoscopic tracking of bodies in motion

Roberto Cipolla and Masanobu Yamamoto*

A fast, highly efficient and robust visual tracking process for multiple moving objects using stereo image sequences taken from a stationary pair of cameras is presented. The algorithm assumes that object motion is restricted to a horizontal plane (e.g. motion of cars on roads or humans walking). Dense stereo image sequences and the visualized locus *method[1,2] (in which each image sequence is first sampled to produce a 2D spatio-temporal cross-section image) are used to ensure temporal correspondence without search. Edge segments in the left and right spatio-temporal images are then matched. Additional stereo matching constraints are derived by using motion and temporal continuity to reduce the number of ambiguous matches. Speed is achieved by only processing a single spatio-temperal cross-section image from each image sequence. The algorithm succeeds in tracking objects in space and time moving against arbitrarily complex backgrounds and in the presence of occlusion, disappearance and reappearance of object features. The output of the algorithm is the 3D position of moving objects as a function of time.*

Keywords: stereo and temporal correspondence, 3D tracking, image sequence analysis

In real dynamic scene analysis the considerable changes in image structure that can occur as a result of object motion (such as the occlusion, disappearance and reappearance of object features) are major obstacles to the successful application of existing 'shape from stereo'[3] and 'structure from motion'[4] techniques to the problem of 3D object tracking from visual data. Monocular image sequence analysis for example, has the following inherent difficulties; the *temporal correspondence* problem in obtaining optical flow locally[4] or matching tokens in discrete views; a *speed-scale ambiguity* which makes it impossible to determine 3D structure and motion in absolute terms for a monocular observer viewing unfamiliar object; and a restriction to rigid body motions which usually require the segmentation of images into parts corresponding to objects with the same motion. These methods often perform poorly with respect to accuracy, sensitivity to noise, and robustness in the face of errors[5,6]. An additional practical problem is that most applications of visual tracking require the real-time processing of large volumes of data. For most existing algorithms this requires special purpose hardware.

Stereo vision can be used independently of motion analysis in a dynamic environment to determine the trajectory of an object in space by taking successive stereo snap shots, determining object locations at each instant and combining these locations into a trajectory[7,8] The most difficult part of the processing concerns the correspondence problem: what to match and how to match it[3].

The correspondence problem is particularly difficult in the presence of occluding boundaries and semi-transparent surfaces such as fences or windows. With dynamic scenes the disappearance and reappearance of image features may make matching and the interpretation of 3D structure instantaneously ambiguous or impossible. An additional drawback in repeating the stereo vision processing at each time instant is that large amounts of data processing (both time and volume) are required. A considerable saving in data processing can be achieved by exploiting knowledge of motion. Ideally, only the edges of objects in motion need be matched. Of greater interest, however, is whether *motion* can interact with stereopsis at the level of matching to help disambiguate false matches. Poggio and Poggio[3] have noted that image motion may be able to aid stereo in the matching process.

MOTION AND STEREO FUSION

Both 'stereo vision' and 'structure from motion' have typically been treated as separate parallel processes. However, each has inherent difficulties, and so it seems logical to attempt to combine them into a single system. Jenkin and Tsotsos[9] and Waxman and Duncan[10] have attempted to unify stereo and motion analysis in a manner which helps to overcome the others shortcomings.

Jenkin and Tsotsos[9] describe a stereo vision system which will track special extracted object feature points in 3D space over time. It uses the 3D interpretation of

Robotics Research Group, Department of Engineering Science, University of Oxford, OX1 3UH, UK
*Computer Vision Section, Electrotechnical Laboratory, Tsukuba 305, Japan

0262–8856/90/010085–06

the feature point velocities to help in the stereo matching process. Even with a sparse feature set (extracted with the Moravec interest operator[7]) the algorithm requires an extensive search in both time and space, and this makes it unsuitable for realtime processing.

Waxman and Duncan[10] have proposed an integrated stereo-motion analysis beginning with the determination of image flow and using correlation between relative image flow (*binocular difference flow*) and stereo disparity locally in establishing correspondence. However, this method suffers from the problems of estimating optical flow, and it requires finely textured smooth surfaces which can be approximated locally as planar. Its application to long real image sequences has not been tested.

As mentioned above, temporal and stereo correspondence are severe problems and limit the usefulness of existing algorithms to applications of visual tracking. An additional practical problem is the need for realtime processing. If, however, we assume a restricted class of motions and process a dense sequence of images so that temporal continuity is guaranteed (as in Bolles' epipolar-plane image analysis[11]) it is possible to greatly simplify (and in a special case avoid) the temporal correspondence problem. If, in addition, we use motion and temporal continuity as additional matching constraints, the stereo correspondence problem can also be considerably simplified.

In this paper we present a fast, highly efficient and robust visual tracking process for multiple moving objects, from stereo image sequences taken from a stationary pair of cameras. Objects moving against arbitrarily complex backgrounds and in the presence of occlusion, disappearance and reappearance of object features are tracked in space and time.

The algorithm presented assumes that object motion is restricted so that its height above a horizontal plane is constant (e.g. motion of cars on roads, humans walking, etc.). We show that dense stereo image sequences and the *visualized locus* method[1,2] (in which each image sequence is first sampled to produce 2D spatio-temporal cross-section images) can be used to ensure temporal correspondence automatically without search. Additional stereo matching constraints are derived by using motion and temporal continuity to reduce the number of ambiguous matches. Computational speed is achieved by only processing a single spatio-temporal cross−section image from each image sequence.

TEMPORAL CORRESPONDENCE: THE VISUALIZED LOCUS

Theory

Consider the perspective projection of an object point onto a planar screen normal to the z-axis. We model the camera as a pin hole with centre at $C(0, H, f,)$, and with focal length f. A point $P(X, Y, Z)$ on an object will be projected onto a point $p(x, y)$ on the image plane such that (see Figure 1a):

$$\frac{X}{x} = \frac{Y-H}{y} = \frac{Z}{f} \tag{1}$$

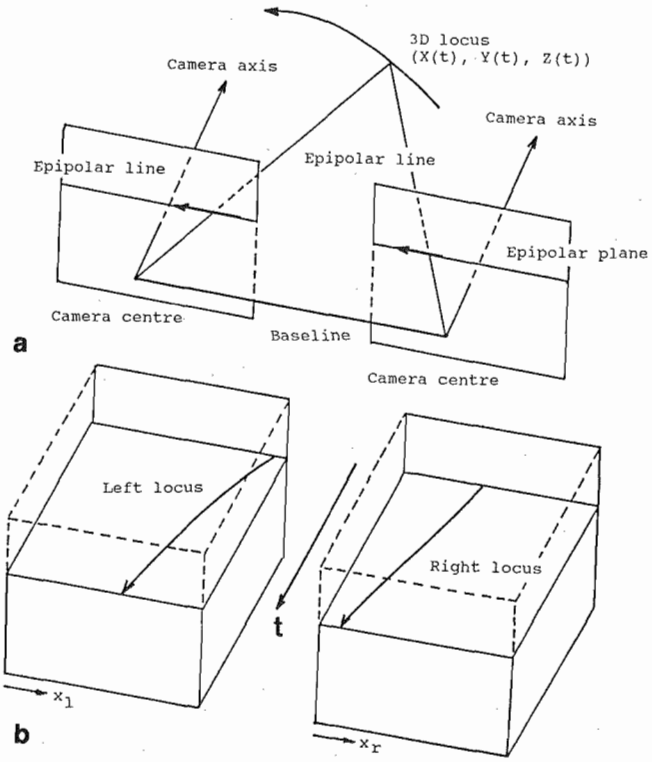

Figure 1. Stereo visualized locus method. (a) camera geometry, (b) 3D spatio-temporal images with cross-sections showing visualized loci

If the object moves the projected point's image position is a function of time, $p(x, y, t)$. The image co-ordinates, x and y, as functions of time are given by:

$$x(t) = \frac{fX(t)}{Z(t)} \tag{2}$$

$$y(t) = \frac{f(Y(t)-H)}{Z(t)} \tag{3}$$

If a sequence of images is taken in rapid succession and piled up sequentially with time, we can construct a 3-dimensional spatio-temporal image with time as the third dimension (see Figure 1b). If temporal continuity from image to image is ensured, the image object point P, $p(x, y, t)$, in *general* draws a 3-dimensional locus in this 3-D image. This locus is referred to as the *visualized locus,* since it is the locus of the projection as a function of time and if the object point P is occluded or goes out of view of the camera the image point p disappears[1,2].

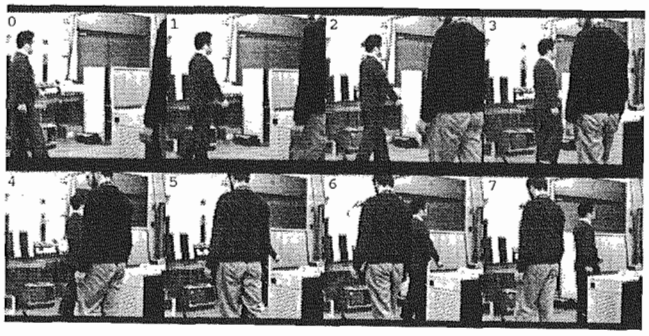

Figure 2 Eight samples (1 second intervals) from left camera image sequence (video rate)

In the case in which the motion of P is constrained to move on a plane $Y(t) = H$ (i.e. at the same height as the optical centre), the visualized locus lies on a plane in the 3D image. For this *special* case, the projection of P is constrained to the same single scan line (raster) of the image at all times. The 2D image synthesized by storing this scan line from each image of the sequence and arranging the set of scan lines sequentially in order of time is a spatio-temporal image. It is an $x-t$ cross-section of the 3D spatio-temporal image. The visualized loci of points at the same height as the camera centre are constrained to this cross-section image and are therefore 2-D loci.

Other researchers have constructed similar images (e.g. see References 11–13). As Bolles *et al.*[11] noted, even though the spatial images which were used to construct it contain complex shapes and intensity changes (see Figures 2 and 3), the spatio-temporal image, as a consequence of smooth motion, is composed of simpler image structures, regions and edge segments (see Figure 4).

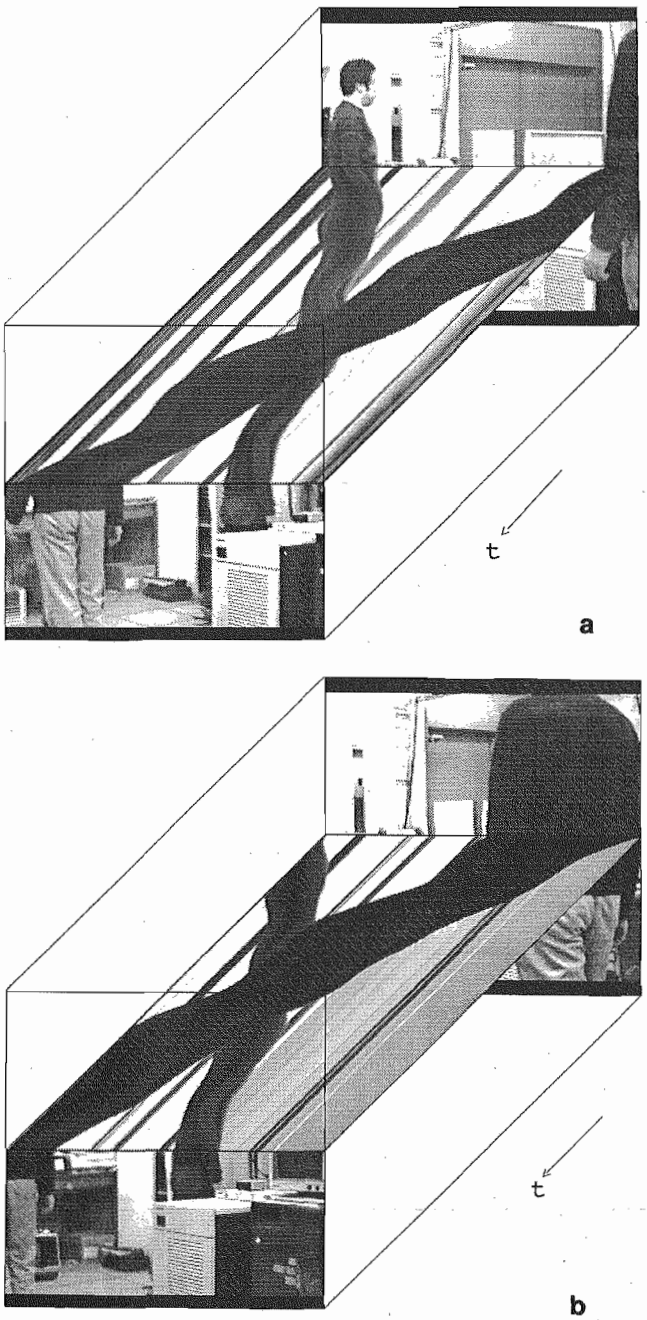

Figure 3. Left and right 3D spatio-temporal images

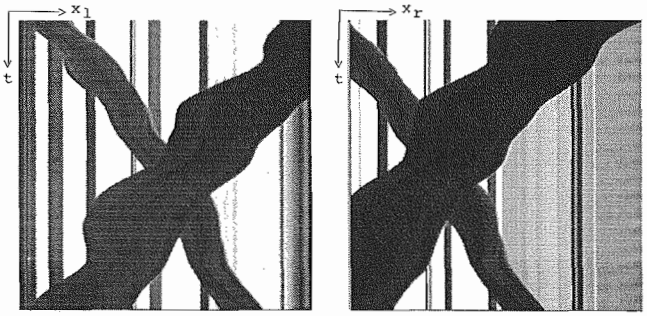

Figure 4. Left and right 2D cross-section spatio-temporal images showing visualized loci of object features at the same height as the camera centres

Higher level properties can also be inferred from the synthesized image. The relationship between loci of points on different objects can be used to determine their relative motions. The disappearance and reappearance of a locus indicates occlusion and by looking at neighbouring loci, the occluding object can be determined.

Unlike Bolles' epipolar-image analysis, however, the spatio-temporal cross-section images used in this method are generated from a pair of stationary cameras with multiple moving objects and non-linear motion. Accurate knowledge of the motion is not required. The original contribution of this paper is the stereo matching of edge segments between spatio-temporal cross-section images generated from the left and right camera image sequences.

EXTRACTION OF VISUALIZED LOCI

The loci of points on the boundaries of regions with high contrast in an image appear as edges in the synthesized spatio-temporal corss-section image. The extraction of these loci thus involve procedures for edge detection and enhancement; the fitting of line and curve segments to these edges; and the merging of segments belonging to the same locus based on temporal continuity. These algorithms are described in Reference 2. The latter step is required because due to occlusion, or camouflage (object has same intensity as background), and noise the extracted loci may be fragmented. Where possible, the fragments are linked by linear interpolation across 'missing' segments. This procedure is based on proximity of the fragment end points and attempting to ensure smooth continuous loci. This latter step is not always possible. It is not essential, but as mentioned below, the disambiguating power of the proposed stereo matching algorithm increases with the length of the extracted loci.

The visualized loci of points on stationary objects appear as straight edges with no time gradient in the synthesized cross-section image. It is therefore very easy to distinguish between the loci of stationary objects and moving objects, regardless of the complexity of the background.

STEREO CORRESPONDENCE

Stereo visualized locus method

If we observe a scene with a stereo pair of cameras and synthesize a spatio-temperal 3D image for both the left and right image sequences there exist two visualized loci corresponding to the same point P: $p_l[x_l(t), y_l(t)]$ and $p_r[x_r(t), y_r(t)]$. If these left and right loci can be correctly matched (stereo correspondence found) the depth as a function of time $Z(t)$, and hence the position ($X(t)$ and $Y(t)$) can be determined for a calibrated camera set by triangulation.

For the special case in which the motion of the body is constrained to a horizontal plane and the cameras are at the same height, the two visualized loci of an object point at the same height as the optical centres are constrained to a single plane in the 3D blocks of data (see Figure 1). The visualized loci ($x_l(t)$ and $x_r(t)$ can then be automatically extracted from the 2D spatio-temporal cross-section image for each image sequence.

The dynamic stereo correspondence problem is then to match the visualized loci in the left spatio-temporal cross-section image with those in the right image. The stereo matching algorithm is an extension of edge-based techniques to edges with time as an additional dimension.

Parallel camera geometry is not necessary. Any camera geometry can be used that ensures scan lines are horizontal (i.e. tilt and convergence are allowed). This is because the epipolar plane for an object point at the height of the optical centres will always be horizontal.

Search for correspondence

For cameras with parallel optical axes, the relative position of corresponding points in the left and right image are geometrically constrained by the *epipolar constraint* and by:

$$x_{pl} \geq x_{pr} \tag{4}$$

which is the *ordering constraint* relative to a point at infinity. This constraint expresses the portion of the epipolar line in the other image on which a potential match may be found. An additional constraint is that of *uniqueness:* each matching primitive should match at most one primitive from the other image.

In the dynamic stereo problem the epipolar constraint is used to generate the left and right spatio-temporal cross-section images — corresponding rasters are epipolar lines for a given time (see Figure 3). The matching primitives are portions of the visualized loci, and we extend the above conventional constraints to encompass *temporal continuity,* i.e. instead of matching at a single time (matching edge pixels along corresponding rasters of the cross-section images), we apply the matching constraint to the visualized locus for all time. The search for corresponding points in left and right images is also considerably reduced by only considering the loci of moving objects.

The search initially finds portions of loci in the left and right images which co−exist (overlap) in time (epipolar constraint); which satisfy the constraint of equation (4) for *all* common times, and which have similar x-direction profiles. Multiple correspondences are reduced by ensuring that each locus in the left image has at least one corresponding candidate for matching in the right image (uniqueness). If ambiguous matches still exist, pairs are chosen which have maximum time overlap and whose disparity changes smoothly with time.

These matching constraints were sufficient for our experiments. Unambiguous correspondence is not guaranteed.

Although it is possible for the matching strategy to be used on static edges there is no advantage over existing stereo algorithms in using the proposed matching algorithm for static edges. In fact, the algorithm's disambiguating power will be poor with static edges, since in this case it is strictly matching along epipolar lines only, and it will be inferior to algorithms which use figural continuity or local support[14].

RESULTS

Eight frame samples of an office scene containing two people moving along the floor are shown in Figure 2. The scene was observed for 8.5 s (512 images (fields) at video rate) with a stereo pair of calibrated TV cameras[15] with a long baseline of 0.340 m and optical centres at a height of 1.2 m above the horizontal ground plane. One person is occluded for part of the observation period. 512×512 spatio-temporal cross-section image are produced from one scan line (average of a swath of seven lines for robustness to interlacing and vertical motion) from each image for both left and right camera image sequences (see Figure 4). This operation can be performed in realtime. For both left and right cross-section images: edge detection (CPU time 2–3 s); segment labelling and curve fitting (CPU time 10 s); and the extraction and description of the visualized loci (CPU time 3 s) are performed on a Sun-3/260 workstation (see Figure 5). The extraction includes merging fragments of loci across 'missing' edges to give long continuous loci (shown as dashed lines). Stereo correspondence and determination of depth was then carried out. The algorithm's output is a plan view of the 3D locus (X-position and depth, Z, Y-position assumed fixed) as a function of time (shown in Figure 6). Solid lines correspond to visible features: dashed lines correspond to features which disappear and

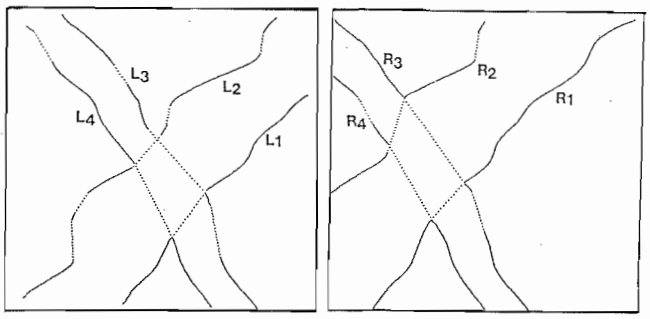

Figure 5. Extracted loci on left and right spatio-temporal cross-section images. Full lines correspond to fragments of the visualized loci, and are linked by dotted lines where the visualized locus disappears due to occlusion, camouflage or noise

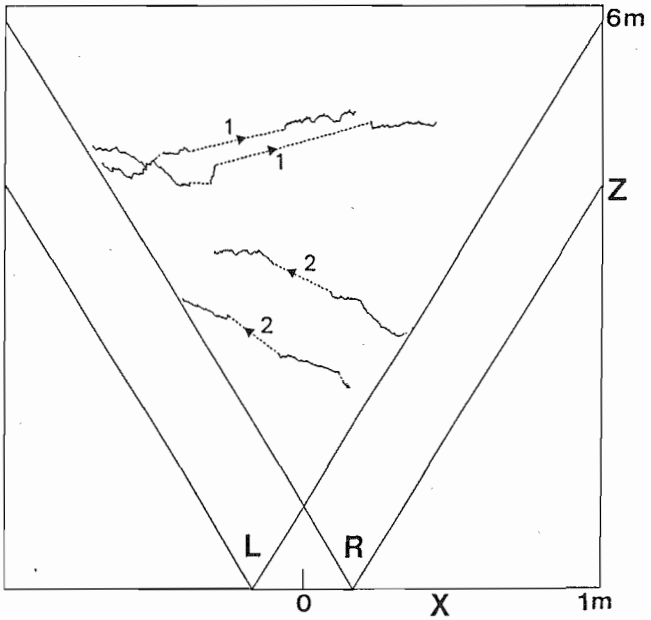

Figure 6. Output of tracking algorithm showing 3D loci of moving objects (plan view) in scene of Figure 2. Two loci are shown for each body. One person is moving from left to right (labelled 1) while the other (labelled 2) is moving from right to left and away from the camera. Solid lines correspond to visible features. Dashed lines correspond to features which are occluded or have the same intensity as the background

reappear as a consequence of occlusion or camouflage or noise.

Additional examples of stereo image sequences of office scenes and the tracking algorithm output are presented in Figure 7 (left spatio-temporal image) and Figure 8 (output) for a scene in which one person overtakes the other, and Figure 9 (first and last frame from left camera) and Figure 10 (output) for a person moving towards the camera in front of a parallel bar fence.

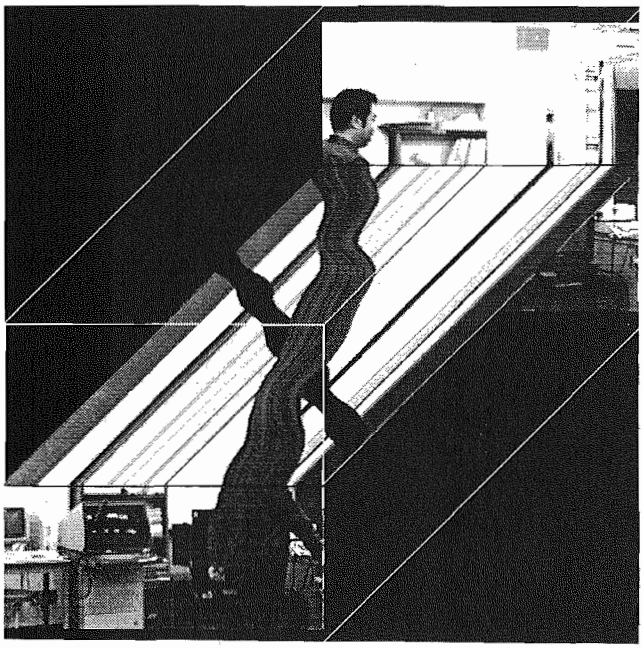

Figure 7. Left 3D spatio-temporal image for scene in which one person overtakes the other

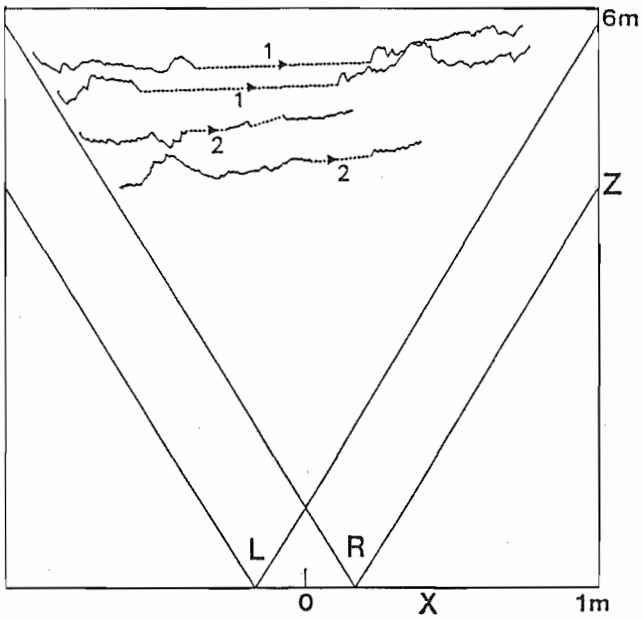

Figure 8. 3D loci of moving bodies (plan view): one person (labelled 1) overtakes the other (2). ◀: progress direction

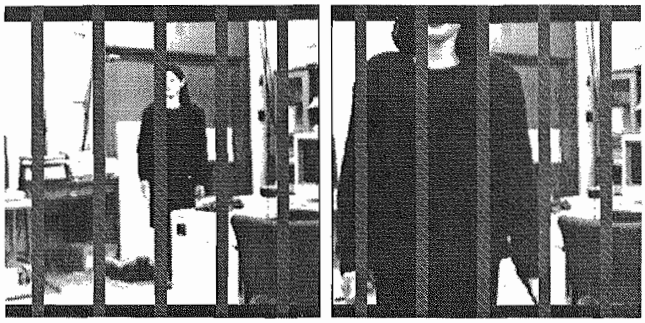

Figure 9. First (left) and last (right) image of left image sequence of a person approaching the cameras behind a fence

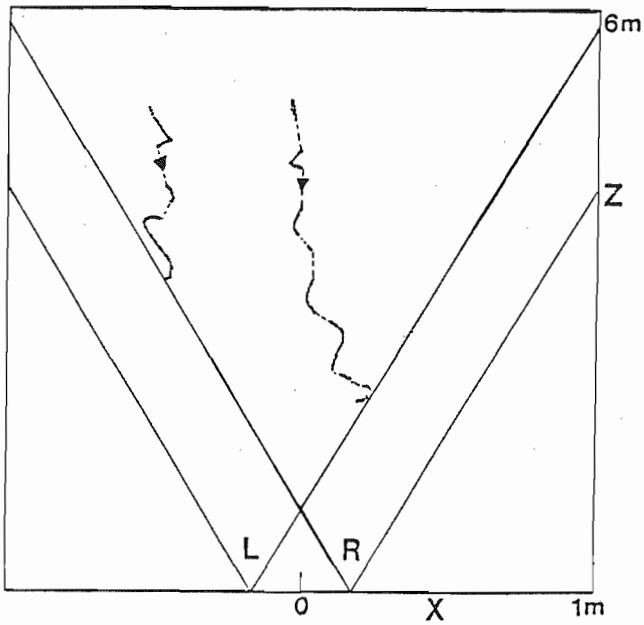

Figure 10. 3D loci of a person approaching camera behind a fence. The two loci shown correspond to the extrema of the left and right arms. They are modulated by an approximately periodic function due to arm motion. ◀: progress direction

In the last example, the object features are continually disappearing and reappearing, and the loci are modulated by an approximately periodic function as a consequence of arm motion modulating the body width as seen by the cameras. At any time instant most existing static stereo matching algorithms would fail to find the correct correspondence or produce ambiguous results, because ordering constraints are not obeyed and because of the lack of local support. The algorithm presented successfully tracks the moving object behind the fence by using temporal continuity to predict where hidden points are based on the object's motion history.

CONCLUSIONS

A method using stereo image sequences to automatically detect and track the 3D motion of objects has been presented. It overcomes the problems of large data storage and processing by processing the single spatio–temporal cross–section image of the image sequence which contains the visualized loci (the locus of image points in time) of object points at the same height or the optical centres of the cameras. This is possible if the motion is restricted to a horizontal plane, and makes unnecessary the need for searching for temporal correspondence. Stereo correspondence between left and right camera images is also greatly simplified by using motion as a cue to suppress the background, and by modifying the stereo matching constraints to encompass temporal continuity.

The algorithm succeeds in tracking objects in space and time moving against arbitrarily complex backgrounds and in the presence of occlusion, disappearance and reappearance of object features. The correspondence techniques used establish matches using only partial information, and make predictions where invisible (hidden) points are given their past motion histories and the motion of their visible neighbours. Stereo matching in the presence of partly transparent objects in the foreground (e.g. fences with parallel bars) has been demonstrated.

The algorithm is robust to calibration errors, image noise and deviations from perfect rigidity (e.g. presence of external boundaries of curved surfaces or the motion of human body). It also works well with large stereo baselines and for long image sequences. It was tested on a variety of scenes to track the 3D motions of humans moving along the ground in various directions relative to the camera.

The motion tracking scheme presented will only correctly track points whose true motion is in the horizontal plane containing the camera optical centres. The resulting algorithms are simple and fast, and can handle object occlusion, disappearance and reappearance. The output is sufficient for determining the number of moving objects in a scene and their general motions.

The methods presented can be extended to general 3D motions. This however, involves a more difficult temporal correspondence problem to extract the 3D visualized loci from the spatio-temporal block of data. The stereo correspondence problem will, however, be simplified.

ACKNOWLEDGEMENTS

Roberto Cipolla is grateful for discussions with Dr Andrew Blake, Professor Mike Brady, Professor Chris Brown and Dr Andrew Zisserman. He acknowledges the support of the IBM UK Scientific Centre. Both authors are indebted to Dr Yoshiaki Shirai and Dr Masaki Oshima and other members of the Computer Vision Section, Electrotechnical Laboratory.

REFERENCES

1 **Yamamoto, M** 'Motion analysis using the visualized locus method *Trans. Info. Process. Soc. Japan.* Vol 22 No 5 (1981) pp 442–449 (in Japanese)

2 **Cipolla, R and Yamamoto, M** *Image Sequence Analysis of Human Motion* Technical Report TR-88-11, Electrotechnical Laboratory, Tsukuba, Japan (1988)

3 **Poggio, G F and Poggio, T** 'The analysis of stereopsis' *Annual Res. Neurosci.* Vol 7 (1984) pp 379–412

4 **Ullman, S** *The Interpretation of Visual Motion* MIT Press, Cambridge, MA, USA (1979)

5 **Barron, J** *A survey of approaches for determining optic flow, environmental layout and egomotion* Technical Report RBCV-TR-84-5, University of Toronto, Canada (1984)

6 **Tsai, R Y and Huang, T S** 'Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces' *IEEE Trans. PAMI* Vol 6 No 1 (1984) pp 13–26

7 **Moravec, H P** 'Visual mapping by a robot rover' *Proc. 6th Int. Joint Conf. Artif. Intell.* (1979) pp 598–600

8 **Ayache, N and Faugeras, O D** 'Building, registration and fusing noisy visual maps' *Proc. 1st Int. Conf. Comput. Vision* London, UK (1987) pp 73–82

9 **Jenkin, M and Tsotsos, J K** 'Applying temporal constraints to the dynamic stereo problem' *Comput. Vision, Graphics & Image Process.* Vol 33 (1986) pp 16–32

10 **Waxman, A M and Duncan, J H** 'Binocular image flows: steps toward stereo-motion fusion' *IEEE Trans. PAMI* Vol 8 No 6 (1986) pp 715–729

11 **Bolles, R C, Baker, H H and Marimont, D H** 'Epipolar-plane image analysis: an approach to determining structure' *Int. J. Comput. Vision* Vol 1 No 1 (1987) pp 7–55

12 **Bridwell, N J and Huang, T S** 'A discrete spatial representation for lateral motion stereo' *Comput. Vision, Graphics & Image Process.* Vol 21 (1983) pp 33–57

13 **Adelson, E H and Bergen, J R** 'Spatio-temporal energy models for the perception of motion', *J. Optical Soc. Am.* Vol A2 No 2 (1985) pp 284–299

14 **Pollard, S B, Mayhew, J F W and Frisby, J P** 'PMF: a stereo correspondence algorithm using a disparity gradient' *Perception* Vol 14 (1985) pp 349–387