Structure-from-Motion under Orthographic Projection

Chris Harris
Plessey Research Roke Manor, Roke Manor, Romsey, Hants, England

Abstract

Structure-from-motion algorithms based on matched point-like features under orthographic projection are explored, for use in analysing image motion from small rigid moving objects. For two-frame analysis, closed-form n-point algorithms are devised that minimise image-plane positional errors. The bas-relief ambiguity is shown to exist for arbitrary object rotations. The algorithm is applied to real images, and good estimates of the projection of the axis of rotation onto the image-plane are obtained.

1 Introduction

Structure-from-motion (SFM) algorithms are used in the analysis of image motion caused by relative three-dimensional (3D) movement between the camera and the (unknown) imaged objects. These algorithms attempt to recover both the 3D structure of the image objects (assumed rigid) and the 3D motion of each object with respect to the camera (or *vice versa*). The SFM algorithms explored in this paper use point image features, extracted independently from each image in the sequence by use of a 'corner' detection algorithm [1], and matched between images forming the sequence [2].

As the imaging mechanism of conventional cameras is perspective projection (ie. cameras behave as if they were 'pin-hole' cameras), most SFM algorithms have been based on perspective projection [3,4]. These algorithms have been found to provide acceptable solutions to the 'ego-motion' problem, where a camera (of relatively wide field-of-view) moves through an otherwise static environment. However, for the perspective SFM algorithms to be well-conditioned, the angle subtended by the viewed object (in the ego-motion problem, the viewed scene) must be large, and the viewed object must span a relatively large range of depths. Thus the perspective SFM algorithms are of little or no practical use for analysing everyday imagery of independently moving objects, such as driven cars and flying aircraft. It is algorithms for the analysis of such imagery that is the concern of this paper.

It is well-known that SFM algorithms are unable to produce an unambiguous solution from visual motion data alone, because of the speed-scale ambiguity. On the analysis of a pair of images, the speed-scale ambiguity dictates that the direction of translation of the camera (relative to the viewed object) may be determined, but not the magnitude of translation, ie. the speed. SFM algorithms are thus carefully constructed to avoid attempting to resolve this ambiguity. However, the current perspective SFM algorithms *do* attempt to solve for all the other motion parameters, no matter how ill-conditioned they may be. A prime example here is the bas-relief ambiguity, where, for example, an indented fronto-parallel surface (ie. viewed head-on) rotates about an axis lying in the surface. The problem of differentiating between a deeply indented surface rotating through a small angle, and a

The author gratefully acknowledges the European Commission who together with the Plessey Company supported this work under the Esprit Initiative project P2502.

shallowly indented surface rotating by a larger angle, is ill-conditioned. Current perspective SFM algorithms applied to such a scene produce diverse and incorrect solutions. This is not to say that such scenes are inherently intractable, but that the SFM algorithm is failing because it attempts to determine the value of ill-conditioned variables. What are needed are algorithms in which the ill-conditioned variables (or combinations of variables) are taken care of explicitly and analytically (just as the speed-scale ambiguity is), leaving only well-conditioned variables to be solved for.

The ill-conditioning that we wish to circumvent occurs for objects subtending a small range of depths, and generally subtending a small angle. In these circumstances, a good approximation to the imaging process is orthographic projection, in which the variation in object depths is assumed to be negligible with respect to the distance of the object from the camera. A further reason for using orthographic projection is that it is mathematically tractable for the analysis of the motion of point-like image features between two images of a sequence [5].

2 Structure-from-Motion Algorithm

Let there be n matches between the two frames, at image locations $\{x_i,y_i\}$ on the first frame, and at $\{x'_i,y'_i\}$ on the second frame. Define a coordinate system with the z-axis aligned along the optical axis, the x and y axes aligned with the image coordinate axes, and the origin at a distance L in front of the centre of projection (the camera pin-hole), so placing the centre of projection at z = -L. Let the i'th point on the moving object be located in 3D at $\mathbf{r}_i = (X_i,Y_i,Z_i)$ at the time of the first frame, and at $\mathbf{r}'_i = (X'_i,Y'_i,Z'_i)$ on the second frame. Below, the object will be assumed to be situated close to the coordinate origin, and be small compared to L. Without loss of generality, decompose the object motion between the two frames as a rotation about the origin, specified by the orthogonal rotation matrix R, followed by a translation **t**. Hence

$$\mathbf{r}'_i = R\,\mathbf{r}_i + \mathbf{t}$$

Perspective projection onto a forward image plane a unit distance from the camera pin-hole gives

$$(x_i ,y_i) = (X_i,Y_i) / (L + Z_i), \qquad (x'_i ,y'_i) = (X'_i,Y'_i) / (L + Z'_i)$$

Substituting gives

$$x'_i = R_{11}x_i + R_{12}y_i + R_{13}z_i + t_x / L + O(L^{-2})$$

and similarly for y'_i. Dropping the $O(L^{-2})$ terms for large L (this is the orthographic limit), and without loss of generality setting L=1, gives

$$x'_i = R_{11}x_i + R_{12}y_i + R_{13}z_i + t_x, \qquad y'_i = R_{21}x_i + R_{22}y_i + R_{23}z_i + t_y$$

Now, for real data, the positions $\{x_i, y_i, x'_i, y'_i\}$ will be contaminated by measurement noise, so that the above equations will not hold true exactly. Assuming isotropic Gaussian noise on the observed image-plane locations, the maximum likelihood solution is found by minimising, E, the sum of the squares of the residuals of the above equations

$$E(R,t_x,t_y,\{z_i\}) = \sum_{i=1}^{n} [\,(R_{11}x_i+R_{12}y_i+R_{13}z_i+t_x-x'_i)^2 + (R_{21}x_i+R_{22}y_i+R_{23}z_i+t_y-y'_i)^2\,]$$

Note that it is the actual residuals of the image-plane locations that are being minimised, and not some other mathematically convenient but less meaningful formulation. To minimise E, first define

$$u_i = R_{11}x_i + R_{12}y_i - x'_i$$

$$v_i = R_{21}x_i + R_{22}y_i - y'_i$$

Thus

$$E = \sum_i [\, (u_i + R_{13}z_i + t_x)^2 + (v_i + R_{23}z_i + t_y)^2 \,]$$

Minimising with respect to z_i gives the optimal depth of each point. Substituting the optimal depths back gives E as a function of the motion variables alone

$$E(R,t_x,t_y) = \sum_i [\, R_{23}u_i - R_{13}v_i + (R_{23}t_x - R_{13}t_y) \,]^2 / [\, R_{13}^2 + R_{23}^2 \,]$$

Note that the x and y translations enter the above equation in a fixed relationship, so that minimising with respect to t_x and to t_y would give identical results. To circumvent this problem, E is minimised with respect to the appropriate linear combination of the translations

$$\partial E/\partial(R_{23}t_x - R_{13}t_y) = 0 \;\Rightarrow\; R_{23}t_x - R_{13}t_y = -\sum_i (R_{23}u_i - R_{13}v_i)$$

Substituting back into E, and simplifying the notation by assuming that mean values have been removed from $x_i, y_i, x'_i, y'_i, u_i$ and v_i, enables E to be written as

$$E(R) = \sum_i [\, R_{23}u_i - R_{13}v_i \,]^2 / [\, R_{13}^2 + R_{23}^2 \,]$$

Without loss of generality, write the rotation matrix, R, as

$$R = \begin{matrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{matrix} \begin{matrix} 1 & 0 & 0 \\ 0 & \cos\eta & -\sin\eta \\ 0 & \sin\eta & \cos\eta \end{matrix} \begin{matrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{matrix}$$

$$= \begin{matrix} \cos\phi\cos\theta + \sin\phi\sin\theta\cos\eta & \cos\phi\sin\theta - \sin\phi\cos\theta\cos\eta & \sin\phi\sin\eta \\ \sin\phi\cos\theta - \cos\phi\sin\theta\cos\eta & \sin\phi\sin\theta + \cos\phi\cos\theta\cos\eta & -\cos\phi\sin\eta \\ -\sin\theta\sin\eta & \cos\theta\sin\eta & \cos\eta \end{matrix}$$

Hence

$$E(\theta,\phi,\eta) = \sum_i [\, (x_i\cos\theta + y_i\sin\theta) - (x'_i\cos\phi + y'_i\sin\phi) \,]^2$$

Note that E is independent of η. This means that there is an irresolvable ambiguity in the rotational motion obtained from two frames in orthographic projection, as reported by Huang [5], and is present for arbitrarily large angles of rotation. This ambiguity is a generalisation of the bas-relief ambiguity found at small angles of rotation, that was discussed in the Introduction. We call the angle η the bas-relief angle. The analytic solution for θ and ϕ is subtle, and ends up by finding the zeros of an 8'th order polynomial. This is performed by use of a standard numerical algorithm, and the solution generated by each real root compared numerically, to see which provides the minimum value of E.

3 Axis of Rotation

The SFM algorithm results in an interpretation of the object motion in which the axis of rotation passes through some point along the optical axis. Were the axis of rotation chosen to pass through some other point (say, because the images were shifted slightly) then an equally satisfactory explanation of the data would result, with unchanged rotation, but with different values for the object translations. The above ambiguity in interpretation can be resolved by seeking a solution in which the object translations are zero; this is appropriate for a scenario where the camera is static, and the object is executing pure rotations.

Let $\mathbf{r} = (x,y,z)$ be any point on the object in the first frame, and $\mathbf{r}' = (x',y',z')$ be the equivalent point on the object in the second frame. These may be actual observed points, or for better conditioning, the centroids of all the matched points may be chosen. Let $\mathbf{c} = (c_x,c_y,c_z)$ be a (3D) point on the axis of rotation. For a motion interpretation with no translation, the motion of the object point about the axis of rotation must be due to rotation alone, hence

$$\mathbf{r}' - \mathbf{c} = R(\mathbf{r} - \mathbf{c})$$

Using the first two equations of this vector equation, and eliminating the term in $z - c_z$, results in the following equation linear in the unknowns c_x and c_y, and so specifies a straight line which is the projection of the axis of rotation onto the image

$$c_x \sin((\theta+\phi)/2) - c_y \cos((\theta+\phi)/2) = (x \cos\theta + y \sin\theta - x' \cos\phi - y' \sin\phi) / (2 \sin((\theta-\phi)/2))$$

That the projection of the rotation axis is independent of the bas-relief angle means that the one-dimensional continuum of possible object rotation axes lie in a plane passing through the camera pin-hole. These conclusions concerning the *orientation* of the rotation axis would still be valid if the object translations had not been chosen to be zero.

4 Affine Projection

The orthographic projection algorithm relies upon the object not significantly approaching or receding from the camera, as this would induce size changes to the image which are not catered for. The addition of such size changes to orthographic projection are called affine projection, and the orthographic algorithm may be modified to cater for them. The equation for the rigid object motion is augmented by a scale (or zoom) factor, s

$$\mathbf{r}'_i = s(R\,\mathbf{r}_i + \mathbf{t})$$

which leads to the following energy term to be minimised

$$E(s,R,t_x,t_y,\{z_i\}) = s^2 \sum_i [(R_{11}x_i+R_{12}y_i+R_{13}z_i+t_x-x'_i/s)^2 + (R_{21}x_i+R_{22}y_i+R_{23}z_i+t_y-y'_i/s)^2]$$

Proceeding as before, the minimisation variables are reduced to s, θ and ϕ

$$E(s,\theta,\phi) = \sum_i [s(x_i \cos\theta + y_i \sin\theta) - (x'_i \cos\phi + y'_i \sin\phi)]^2$$

Defining $p = s\cos\theta$ and $q = s\sin\theta$ enables E to be easily minimised using no more than an arctangent.

5 Results

The inability to resolve the bas-relief ambiguity makes the results of this algorithm difficult to appreciate, as both the rotation and the structure of the object depend upon the bas-relief angle. The SFM algorithms were applied to a sequence of 16 real images of a toy truck on a turntable. The images are 128 pixels square, and the truck subtends an angle of about 5° from the camera. Between each frame of the sequence, the truck was rotated by 10° about an axis passing through the centre of the turntable, and oriented some 3° clockwise of the vertical. Thus the true projection of the axis of rotation is a nearly vertical line in the image, about one third of the image width from the right-hand edge of the image.From each image feature-points were extracted using a corner detector [1], from 20 to 30 being extracted from each image; these are indicated by the black crosses in the Figures. The feature-points were matched by hand for expediency.

The Figures show the projection of the calculated axis of rotation for the analysis of various image pairs. In Figure 1, successive pairs of images are analysed, corresponding to a rotation of 10° between images. The flow-vector of each matched point is shown as a short white line, which terminates at the location of the feature-point in the later of the image pairs. The projection of the calculated axis of rotation for the orthographic algorithm is indicated by the white line spanning the image, and that for the affine algorithm by the black line (it is sometimes wholly or partly obscured by the white line). Figures 2 and 3 show the results for 20° and 40° rotation respectively, and in Figure 4 are shown four results each for 60°, 80°, 100° and 120° rotations in successive rows of the image.

For even the 10° rotations, the SFM algorithms perform well, and for larger angles of rotation the rotation axis is increasingly accurately positioned. Both algorithms produce quite similar results, mainly because there was no significant zooming of the object (the calculated zooms were all close to and consistent with a value of unity). The accuracy of the results increase less than proportionately with the angle of rotation. This is due to the number of matches decreasing as the angle increased, and due to the inconsistency in objective positioning of the feature-points with large object movement.

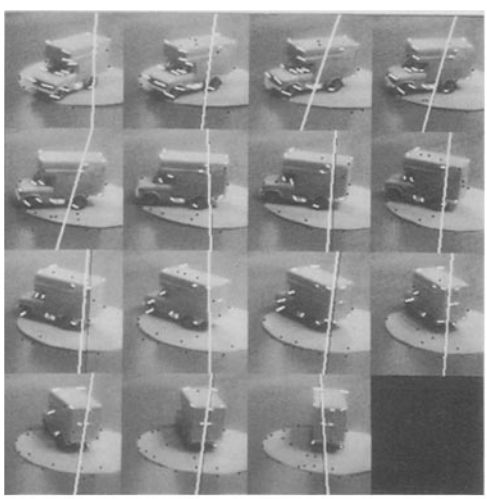

Figure 1. 10° rotation.

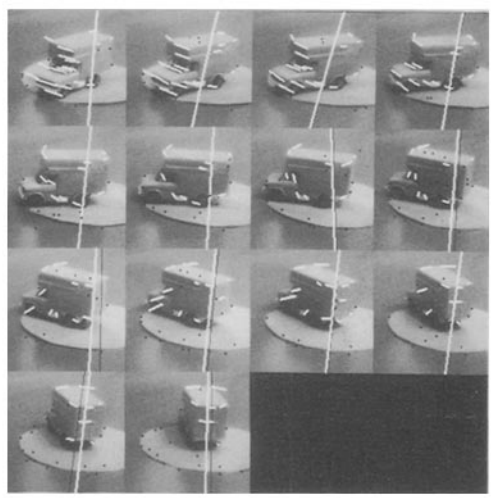

Figure 2. 20° rotation.

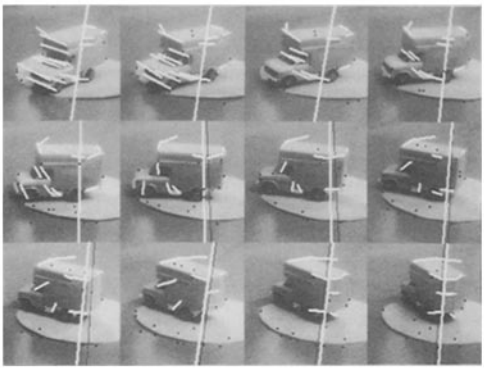

Figure 3. 40° rotation.

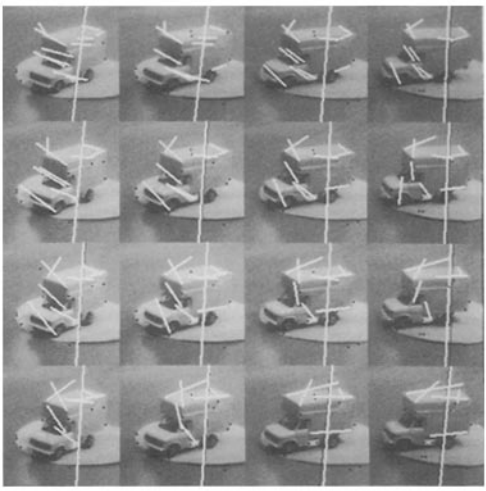

Figure 4. 60°, 80°, 100° and 120° rotation.

6 Conclusions

The two-frame orthographic and affine SFM algorithm have been demonstrated to be well-founded (since they minimise image-plane positional errors), closed-form in implementation, and produce good and well-conditioned results. The existence of the bas-relief ambiguity means that there is an ambiguity in interpreting both the rotational motion and the structure, resulting in only a limited number of readily interpretable outputs, such as the projection of the axis of rotation. To resolve the bas-relief ambiguity, matches from three frames in a sequence are needed, and algorithms for analysing such data are currently under development.

7 References

1 Harris, CG and MJ Stephens, *A Combined Corner and Edge Detector*, Proc. 4th Alvey Vision Conference (1988), pp.147-152.

2 Harris, CG and JM Pike, *3D Positional Integration from Image Sequences*, Proc. 3rd Alvey Vision Conference (1987), pp. 233-236.

3 Harris, CG, *Determination of Ego-Motion from Matched Points*, Proc. 3rd Alvey Vision Conference (1987), pp.189-192.

4 Faugeras, O, F Lustman and G Toscani, *Motion and Structure from Motion from Point and Line Matches*, Proceedings IEEE International Conference on Computer Vision, pp. 25-34, 1987.

5 Huang, TS and CH Lee, *Motion and Structure from Orthographic Projection*, IEEE Trans. PAMI, Vol 11, No 5, May 1989, pp. 536-540.