Applying Visual Processing to GPS Mapping of Trackside Structures

D. C. Nicholls and D. W. Murray
Department of Engineering Science, University of Oxford,
Parks Road, Oxford, OX1 3PJ, U.K.
dwm@robots.ox.ac.uk

Abstract

This paper describes an application of pointwise structure from known motion to assist in the construction of roadside and trackside maps. A single camera is used to record the view from the front of the vehicle, and positional information from an onboard GPS receiver is used to provide the location and orientation of the vehicle. After calibration from known structures, the transformation from world coordinates to image coordinates can be specified for each frame, and multiple views of the same feature, allow the world coordinates of the features to be recovered. Objects must be selected and labelled by hand, but thereafter tracking, recovery and mapping are automated. Examples are given of the recovery and logging of structures along a railway line.

1 Introduction

Maps of roads, railways and waterways can be routinely created and revised nowadays using data from GPS receivers [7] carried along the route (eg [3, 1]). Using differential[1] and other signal analysis techniques, commercial GPS receivers of increasing sophistication and cost quote positional accuracies in the range ± 5m to ± 0.1m, figures which can be improved upon if the motion of and noise in the receiver is modelled within a filter [6], and if other inertial guidance data is available to patch over loss of GPS signal. Such mapping is particularly useful in three areas: (i) remote regions with such sparse detail that aerial surveying is uneconomic; (ii) regions where climate or man's activity mean that roads move; and (iii) in more populous regions where detailed visual logging of wayside structures is required. It is the last application that we are concerned with here, focussing on the application to railways and trackside furniture.

For this task, the typical methodology used commercially is to direct two cameras to the left and right of the moving vehicle (Figure 1a) so that a particular pixel column in the camera is perpendicular to the road or track. Each video frame is timestamped by the GPS, so that frame and location can be married up later. The two videos are viewed offline by eye, and the moment an object passes the fiduciary column are logged. The resulting description of an object is then "leftside" or "rightside" at the current position of the GPS. The weaknesses are obvious enough: two cameras are required; the visual data are all but thrown away; and, because the cameras are pointing laterally, the viewer can be surprised by near objects moving rapidly through the field of view. Given the sophistication of the GPS system (involving not least a few tens of satellites!), just a little more care with the visual data seems warranted.

[1] which provide independence from Selective Availability imposed on Navstar GPS by the US Department of Defense

BMVC 1998 doi:10.5244/C.12.84

One system that does take more care is the GPSVan developed over several years at the Ohio State University [2]. This system uses stereo cameras, capturing digital stereo pairs every 30m or so. The stereo cameras use a baseline of order 2m and the depths of particular objects are found from a single pair of images, given the known position of the two cameras.

In this paper we recover structure from a monocular image stream, given the known motion and hence motion of the single camera. The data is recorded on analogue tape from a standard video camera and digitized later, making the accumulation of data less specialized than in [2]. A map reference together with height are recovered for each selected object by tracking it in a single front facing camera (Figure 1b). The problem of surprise in the entirely manual process is removed because the objects of interest can be seen approaching but, as shown later, because we can guarantee that they will be passed, they do not have to be tracked until quite close to the vehicle.

Section 2 outlines the basis of the method, Section 3 describes the implementation, and Section 4 shows mapping results obtained from a railway line, comparing output with that from the Ordnance Survey. Although there is little novel in the visual processing per se, the paper might serve as a reminder to the vision researcher that a modest financial investment in a GPS receiver makes structure from known motion feasible outside the laboratory and away from the robot arm.

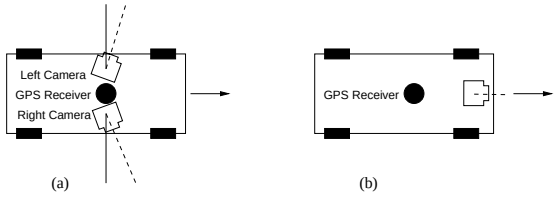

Figure 1: The typical existing method of visual logging (a) uses two cameras directed laterally, and objects are marked when passing the perpendicular. In the present method (b) a single forward-facing camera is used and objects are tracked over several frames.

2 Algorithm

The aim in map generation is to find the positional reference for individual objects, but then only to represent the object symbolically. There is thus no requirement to recover dense structure, but there is a need to recognize and select objects. In this work recognition and selection is performed by the operator. For busy and varied street scenes, operator intervention seems inevitable, but in the less cluttered and more stylized trackside environments explored here it may indeed be possible to automate at least part of this stage.

Once selected, a point on the object o is tracked over images i until it exits the field of view giving rise to n image measurements $(x, y)_{oi}$. Using homogeneous coordinates, the projection into the image is

$$\mathbf{x}_{oi} = \begin{pmatrix} \lambda x \\ \lambda y \\ \lambda \end{pmatrix}_{oi} = [\mathsf{K}][\mathsf{P}]\mathbf{X}_{oi}^{C}$$

where $[\mathsf{K}]$ holds the camera intrinsic parameters, determined by calibration, and $[\mathsf{P}]$ is the 3×4 projection matrix $[\mathsf{I}|\mathbf{0}]$ [4].

The object coordinate $\mathbf{X}_{oi}^{C}$ in the camera frame C is related to the desired static world coordinate $\mathbf{X}_{o}^{W}$ by the concatentation of two transformations, a variable one between the world frame W and GPS reference frame G within the vehicle, and a fixed one between GPS and camera frames (Figure 2(a):

$$\mathbf{X}_{oi}^{C} = \begin{bmatrix} [\mathsf{R}]_{G}^{C} & -[\mathsf{R}]_{G}^{C}\mathbf{t}_{CG} \\ \mathbf{0}^{\top} & 1 \end{bmatrix} \begin{bmatrix} [\mathsf{R}]_{W}^{G} & -[\mathsf{R}]_{W}^{G}\mathbf{t}_{GW} \\ \mathbf{0}^{\top} & 1 \end{bmatrix}_{i} \mathbf{X}_{o}^{W} .$$

The GPS frame has its X- and Y- axes along the longitudinal and transverse axes of the vehicle, so that X is a tangent and Y normal to the space curve following by the vehicle. Z is the bi-tangent, polarized to have an upwards component (Figure 2a). The variable transformation thus involves the translation $\mathbf{t}_{GW}$, the position of the GPS receiver in the world frame at image i, and a rotation which is decomposed into a combination of heading elevation and roll $[\mathsf{R}] = [\rho_i][\epsilon_i][\phi_i]$, in the usual order shown in Figure 2(b). In practice, GPS delivers rather poor altitude values, but as the rate of change of altitude is small it is sufficient to assume the space curve lies in a plane parallel to the world's X-Y plane. The the variable elevation ϵ_i and roll ρ_i are set to zero, and the variable heading ϕ_i found by fitting a tangent to the GPS X, Y measurements.

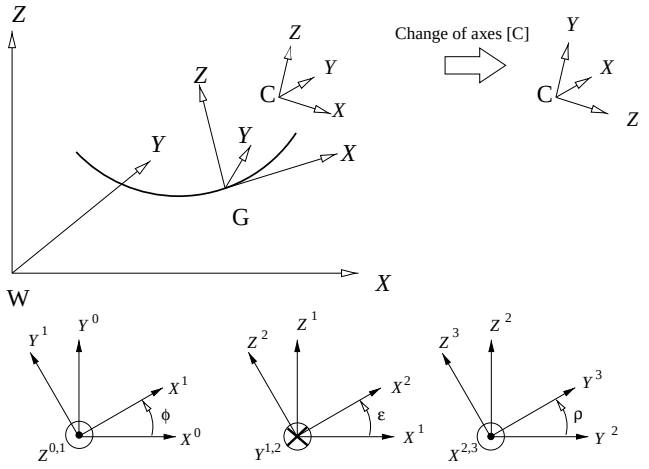

Figure 2: (a) The world, GPS and camera frames. (b) The rotation is built from a change of heading, elevation and roll.

The fixed transformation describes the offset translation $\mathbf{t}_{CG}$ and offset rotation of the camera from the GPS frame. Again the rotation is decomposed into heading, elevation and roll, but a final change of axes is required to use the convention that the camera's optic axis is along Z. $[\mathsf{R}] = [\mathsf{C}][\rho][\epsilon][\phi]$ where all C_{pq} are zero except $C_{12} = C_{23} = C_{31} = C_{44} = 1$. The offsets are found by calibration as described in section 3.

Combining the various transformations we have

$$\begin{pmatrix} \lambda x \\ \lambda y \\ \lambda \end{pmatrix}_{oi} = [\mathsf{M}]_i \mathbf{X}_{o}^{W} = [\mathsf{M}]_i \begin{pmatrix} \mathbf{X'}_{o}^{W} \\ 1 \end{pmatrix}$$

where $[\mathsf{M}]$ is a known 3×4 matrix. Given observations over frames $i = 1...n$ the $2n \times 3$ matrix $[\mathsf{A}]$ and $2n \times 1$ vector $\mathbf{b}$ are constructed

$$\begin{pmatrix} \vdots & \vdots & \vdots \\ (M_{31}x_i - M_{11}) & (M_{32}x_i - M_{12}) & (M_{33}x_i - M_{13}) \\ (M_{31}y_i - M_{21}) & (M_{32}y_i - M_{22}) & (M_{33}y_i - M_{23}) \\ \vdots & \vdots & \vdots \end{pmatrix} \mathbf{X'}_o^W = \begin{pmatrix} \vdots \\ (M_{14} - M_{34}x_i) \\ (M_{24} - M_{34}y_i) \\ \vdots \end{pmatrix}$$

and the system $[\mathsf{A}]\mathbf{X'}_o^W = \mathbf{b}$ solved in the least-squares sense using the pseudo-inverse or, preferably, by singular value decomposition. Care is taken to centre the data using a linear transformation before solving.

The solution from this linear method is then used as the starting point for a non-linear optimization (eg [4]) which finds the $\mathbf{X}$ which minimizes the sum of the squares of the distances of the predicted projections $\mathbf{x}(\mathbf{X}_o^W, [\mathsf{M}]_i)$ from measured image positions

$$\min_{\mathbf{X}_o^W} \sum_{i=1}^{n} (\mathbf{x}_{oi} - \mathbf{x}(\mathbf{X}_o^W, [\mathsf{M}]_i))^2 \quad .$$

3 Implementation issues

3.1 Calibration

The method adopted requires that the camera's intrinsic parameters and extrinsic parameters with respect to the GPS frame be known. Rather than using laboratory techniques to calibrate, we have used large outdoor structures with known geometry combined with the GPS data.

3.1.1 Camera intrinsics

Measurements were made of the image width and height of a bridge of known scene dimensions (Figure 3a). Taking the width as an example, starting from some unknown distance D_0 from the bridge of width W, the image width w was measured as a function of distance z moved forward, a value provided by the GPS independently of offsets between the GPS head and camera. Now

$$\frac{w}{f_x} = \frac{W}{D_0 - z} \quad \text{or} \quad \frac{1}{w} = \frac{D_0}{f_x W} - \frac{1}{f_x W} z \; .$$

Figure 3(b) shows the results of straight line of $1/w$ vs z moving a distance of some 160m towards the bridge. From the slope of the graph and using $W = 8.9(2)$m, we recover $f_x = 614(14)$pixel. From the z-intercept of the graph we find $D_0 = 183(1)$m, a figure we shall return to below. Similar f_y was found to be $623(16)$ pixels, confirming the expected aspect ratio of near unity. Henceforth we take $f = f_x = f_y$.

The principal point was not measured and is taken to be at the image centre $(u_0, v_0) = (192, 144)$.

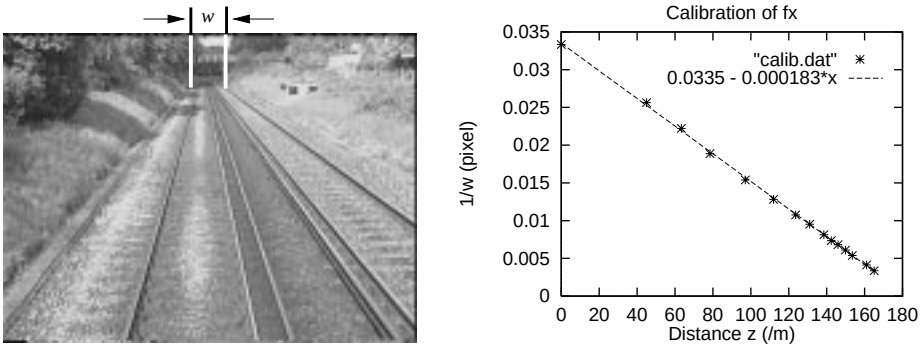

Figure 3: Measured value of $1/w$ against z, and the fitted straight line. The slope is $1/f_x W = 1.83 \times 10^{-4}\text{pix}^{-1}.\text{m}^{-1}$ and the z-intercept is $D_o = 183\text{m}$.

3.1.2 Extrinsic offsets from the GPS frame

(Figure 3(a) was also used to determine the fixed transformation between camera and GPS frames. Because the tracks are parallel to the GPS X-axis, their vanishing point defines the direction of this axis in the image provide a measure of the elevation and heading offsets. By intersecting straight line fits to the tracks, $\mathbf{x}_V = (184.1, 31.9)$ and, taking ρ and ϕ to be negligible, $\epsilon = \tan^{-1}\left[(y_V - v_0)/f_y\right] = -0.18\text{rad}$. The small ρ and ϕ are highly correlated, but each is of maximum value 0.01rad and can be safely neglected within errors.

In the camera's coordinate system (*after* changing the axes using $[\mathsf{C}]$), a point (X, Y, Z) on the ground plane satisfies

$$Y \cos\epsilon + Z \sin\epsilon = -t_{ZCG}$$

where t_{ZCG} is the height of the camera about the ground plane, as shown in Figure 4. Consider now an individual rail lying in the ground plane at $X = d$. Under perspective projection, the imaged rail is the straight line

$$(y - v_0) = +f \tan\epsilon - (x - u_0)\frac{t_{ZCG}}{d\cos\epsilon}$$

with slope $S = -t_{ZCG}/d\cos\epsilon$. If we fit lines to both left and right rails,

$$\frac{1}{S_R} - \frac{1}{S_L} = \frac{\cos\epsilon}{t_{ZCG}}(d_L - d_R) = \frac{\cos\epsilon}{t_{ZCG}}G \ .$$

where G is the known gauge. Using the UK gauge [5] of $G = 1.435\text{m}$ (or in decent units, 4ft $8\frac{1}{2}$in), from the straight line fits to the rails $t_{ZCG} = \cos\epsilon/0.3286\text{m}$ and, using the value of ϵ, we find the height of the camera as $t_{ZCG} = 2.99(5)\text{m}$, a figure entirely compatible with the height of the front window of the driving car of the train.

The camera was at $t_{XCG} = 10\text{m}$, ie 10m in front of the GPS receiver in these experiments, a measurement verified in two ways. First recall that from Figure 3(b) the initial visual distance to the bridge was $D_0 = 183\text{m}$. Now the initial GPS reading was (404606.5, 91569.4)m and the OS map position of the bridge is (404414, 91553)m, giving a difference of 193m. A second verification was made by finding the image and

hence GPS reading where the bottom of the image is just aligned with the front of the bridge. The GPS reading was (404431, 91554)m compared the map position of (404414, 91553)m, placing the GPS 17m from the bridge. From this we subtract the the distance p to the nearest visible point on the ground (Figure 4). Now

$$\tan(-\epsilon + \alpha) = t_{ZCG}/p\,,$$

where the half field of view of the camera α in the vertical direction is found as $\alpha = \tan^{-1}(v_0/f)$. From this we find $p = 6.9m$, verifying that the GPS is some $17 - 6.9 \approx 10$m further back than the camera.

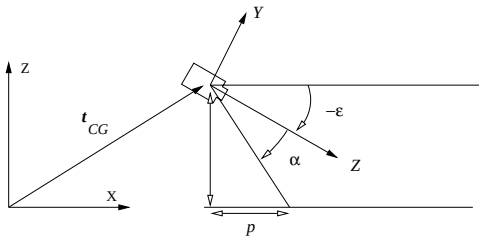

Figure 4: Transformation between the GPS and Camera Frame. The height of the camera above the ground plane is t_{ZCG} and the nearest visible point a distance p away along the ground plane.

3.2 Object selection and tracking

Object selection is carried out by the operator. The object is described as one of a number of structures, such as bridges, huts, signs, switchboxes, buildings, platform starts and ends.

An 21×21 pixel mask centred at the selected point is stored, and tracking is performed from frame to frame using normalized grey-level correlation. A correlation maximum is found to the nearest integer, and quadratic fitting to the correlation surface around the maximum is used to obtain sub-pixel acuity. A new mask is then generated by interpolation for matching to the next image. With two or fewer matches search for matches is restricted to a band around the epipolar line, but with more matches the 3D position is well enough determined to restrict search to a small window about the predicted projection.

The recomputation of the correlation mask is essential to reduce feature drift. This problem is further mitigated by limiting the start of tracking until the object is quite close to the camera. The underlying rationale is that because the object is at the side of the track, its *last* observation can be assured to be close to the edge of the image, and thus we can determine an optimum position for its *first* observation, a position which is not so much further from the camera.

Consider Figure 5, which shows the geometry for two viewing positions of an object laterally displaced from the camera[2] at $X = W$. It is obvious that the best localization will always be obtained if one of the views is the closest possible, where the object is imaged at the extreme pixel w, with a position error of $\pm\beta$. The bounds of uncertainty

[2] Again using the coordinate system *after* application of $[\mathsf{C}]$.

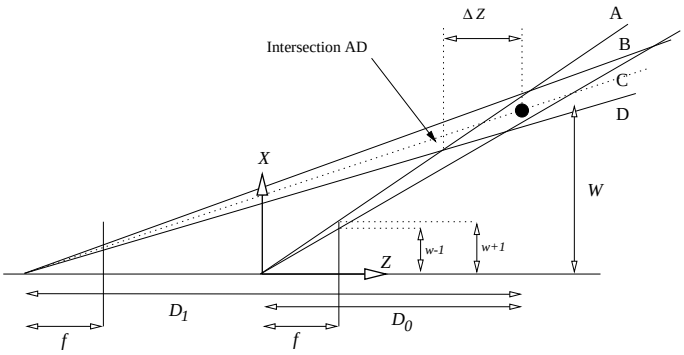

Figure 5: Viewing geometry for an object which is at the edge of the image at distance D_0. Distance D_1 is found to minimum the depth uncertainty ΔZ.

are delineated by lines A and B. The second view is taken from a distance D_1 and the $\pm\beta$ lines are C and D, which with A and B define a quadrilateral of uncertainty. The narrow dimension of the error quadrilateral is all but fixed, and reducing uncertainty is equivalent to making the distance ΔZ a minimum. A little manipulation shows that the intersection of A and D has

$$\Delta Z = \beta \frac{(D_0 + D_1) D_0 D_1}{D_1 f W + 2\beta D_0 D_1 - D_0 f W} ,$$

and differentiating with respect to D_1 and setting to zero yields a quadratic for the optimum D_1:

$$(1 + 2\beta D_0 / f W)\, D_1^2 - 2 D_0 D_1 - D_0^2 = 0 .$$

Only one solution is valid,

$$D_1 = D_0 \left(\frac{1 + \sqrt{2(1 + \beta D_0 / f W)}}{1 + 2\beta D_0 / f W} \right) \approx D_0 (1 + \sqrt{2}) .$$

Thus measurements taken a substantially greater distances than D_1 are of little value. The analysis also gives a feeling for the minimum likely error of

$$\Delta Z = \frac{\beta}{w} D_0 (1 + \sqrt{2})^2 \quad \text{and} \quad \Delta X \approx \frac{\beta}{f} D_0 \frac{(2 + \sqrt{2})}{2} .$$

4 Experiments

Video and GPS data were recorded at 25Hz during a hour long journey along a railway line through Poole. Each video frame had a VITC time code written at the top of the image, allowing video and GPS data frames to be rematched during offline analysis. The experiments shown here use images sampled at 2Hz (actually at alternating intervals of 480 and 520ms).

4.1 Tests on individual structures

We compare the output from the method with two structures whose position is available from a 1:1250 scale OS map, and where the considerable care has been taken with the matching.

Figure 6(top) shows four images taken during the approach to a milepost whose positions is read from the OS map as (404495, 91557)m. Table 1(a) shows the recovered world coordinates as the number n of data used increases. The final value of (404494.1, 91555.9, 0.3)m is agrees with the OS coordinate to within errors.

Figure 6(bottom) shows four images taken during the approach to a bridge whose left pillar is read from the OS map as (404414, 91549)m. The recovered X, Y, Z shown in the table are in agreement.

Finally, a distinctive patch of ballast was tracked. (The images are not printed as the loss of resolution makes the ballast somewhat less than distinctive.) The X and Y values in Table 1 are of little interest, but it will be seen that the Z is zero within the error tolerance.

Figure 6: Four frames covering the last 2 second as the train approaches a mile marker (top) and bridge (bottom).

4.2 Extended test with assorted trackside clutter

We now give examples of the map output. In practice the output is destined for a GIS system, but here we generate a simple graphical output showing track, bridges and so on. It is worth noting first that the train was moving in the westwards direction and so location is based on the eastwards facing parts of the structures, and second that bridge symbols have been oriented perpendicular to the track direction (it would be desirable of course to make measurements of each pillar).

The first example shows a section about 800m east of Parkstone station in Poole and compare it with the OS map of that area.

The second example in Figure 8 shows the mapping of Parkstone Station itself. As with the first example, only the text has been added by hand. Note that the train is travelling on the left hand, more southerly, track, and this is recovered in the map.

MileMk Frames	X or E (m)	Y or N (m)	Z (m)
19-20	404493.9	91555.8	0.4
19-21	404493.7	91555.8	0.3
19-22	404494.5	91556.0	0.4
19-23	404494.3	91555.9	0.4
19-24	404494.1	91555.9	0.3

Bridge Frames	X or E (m)	Y or N (m)	Z (m)
42-43	404417.6	91550.0	1.7
42-44	404414.7	91549.4	1.6
42-45	404414.6	91549.3	1.6
42-46	404413.8	91549.1	1.5

Ballast Frames	X or E (m)	Y or N (m)	Z (m)
7-8	404540.9	91564.5	0.2
7-9	404538.7	91564.3	-0.0
7-10	404538.6	91564.4	-0.0
7-11	404538.5	91564.4	-0.0
7-12	404537.9	91564.4	-0.1

Table 1: Recovered world coordinates of the mile marker and bridge (shown in the previous figure), and of a patch of ballast on the track.

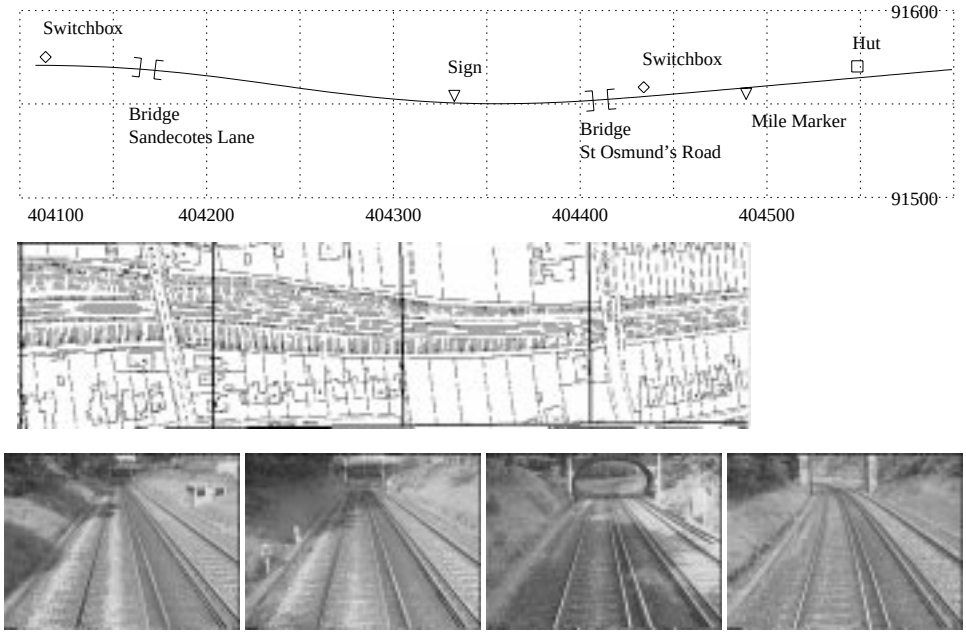

Figure 7: The recovery of a section of track and its comparison with the same section of OS map. Below are some of the images used. The train was travelling in a westerly direction (ie right to left).

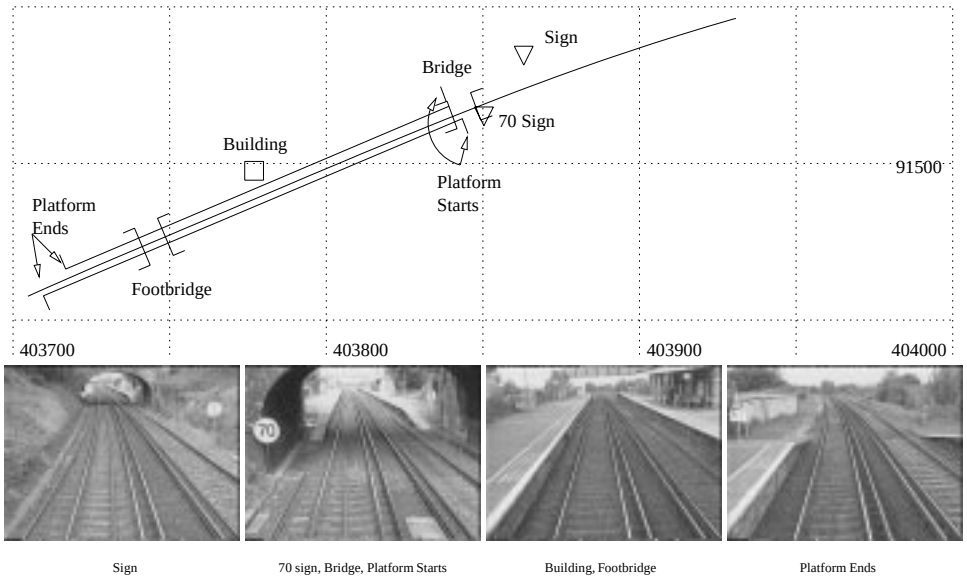

Figure 8: Parkstone Station, Poole. Only the text and arrows are added manually. Below are some of the images used taken from the train which was travelling in a south westerly direction (ie top-right to bottom-left).

Acknowledgements

The authors are grateful to Geographix Ltd for supplying the video and GPS data using a GPS receiver from Guidance Control Systems Ltd. This work was supported by undergraduate project funds from the Department of Engineering Science.

References

[1] D. Bochenek, R.S. Davis, L.B. Swift, and J.M. White. Mapping automation using GPS, GIS and programming. In *Proc AM/FM International, Nashville TN, 23-26 Mar 1997*, pages 775–781. Aurora Co, 1997.

[2] J.D Bossler and C.K. Toth. Accuracies obtained by the GPSVan. In *Proc GIS/LIS '95 Annual Conference, Nasville, TN, 14-16 Nov 1995*, pages 70–77. American Soc Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, 1995.

[3] H-J. Euler, C.D. Hill, and U. Miller. Real time precise GPS for railroad mapping. In *Proc IEEE Position Location and Navigation Symposium, Atlanta GA, 22-26 April 1996*, pages 437–443, New York, 1996. IEEE.

[4] O.D. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, MA, 1993.

[5] C. Lockwood. *Railway Permanent Way*, volume II of *Kempe's Engineers' Yearbook*, chapter L9. Morgan Grampian Book Publishing, London, 1990. ISBN 0 86213 096 4.

[6] E.M. Nebot, H. Durrant-Whyte, and S. Scheding. Kalman filtering design techniques for aided GPS land navigation ppplications. In *Proc 1st Australian Data Fusion Symposium, Adelaide, Australia, 21-22 Nov 1996*, pages 83–88. IEEE, New York, 1996.

[7] B.W. Parkinson and J.J. Spilker, editors. *Global Positioning System: Theory and Applications*. Progress in Astronautics and Aeronautics. American Institute of Aeronautics and Astronautics, Washington, DC, 1996.