Vehicle Shape Approximation from Motion for Visual Traffic Surveillance

George S.K. Fung, Nelson H.C. Yung, and Grantham K.H. Pang
Laboratory for Intelligent Transportation Systems Research
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong SAR
E-mail: skfung, nyung, gpang@eee.hku.hk

Abstract—In this paper, a vehicle shape approximation method based on the vehicle motion in a typical traffic image sequence is proposed. In the proposed method, instead of using the 2D image data directly, the intrinsic 3D data is estimated in a monocular image sequence. Given the binary vehicle mask and the camera parameters, the vehicle shape is estimated by the four stages shape approximation method. These stages include feature point extraction, feature point motion estimation between two consecutive frames, feature point height estimation from motion vector, and the 3D shape estimation based on the feature point height. We have tested our method using real world traffic image sequence and the vehicle height profile and dimensions are estimated to be reasonably close to the actual dimensions.

I. INTRODUCTION

In Visual Traffic Surveillance (VTS), there has been active research in vehicle detection, tracking, counting, and classification[1]. To accomplish these difficult tasks under real world environment, it appears necessary to use some advanced video processing method to explore the intrinsic information of the vehicle hidden in the image sequence. Success in this relative low-level process will guarantee more promising results in the high-level process, such as matching with vehicle generic model.

Recently, a visual-based dimension estimation method for vehicle type classification was proposed [2]. The simple 3D cuboid model was employed to fit any vehicle types. The results showed that the modeling method could effectively estimate the vehicle dimensions, including length, width and height of the vehicle. The estimation accuracy was sufficient for general vehicle type classification. Instead of simple cuboid model, a generic and complex vehicle model based approach was proposed in [3] to track vehicle in monocular image sequence of road traffic scenes,. They used a parameterized 3-D generic model to represent the various types of vehicles moving in the traffic scene. Firstly, the edge segments of the parameterized 3-D polyhedral model were projected from the 3-D scene back into the 2-D image. Then, the matching between 3-D model data and 2-D image data was performed on these 2-D edge segments using the Mahalanobis distance between the attributes of the line segments. Finally, the set with the best correspondence between 3-D model edge segments and 2-D image edge segments was found by using an iterative approach. The matching process of the image data with the vehicle model was carried in the 2-D image environment only, while the intrinsic 3-D information of the vehicle over image sequence was not utilized. This observation motivates us to explore the 3-D information from the vehicle motion.

There are numerous methods to estimate object structure or motion from feature correspondence as discussed in [4]. A simple vehicle height approximation by tracking the feature points over the image sequence was proposed in [5]. In their method, the feature points belonging to the same vehicle were grouped by the common motion constraint, such that feature points that were seen as moving rapidly together would be grouped together into a single vehicle. Over the image sequence, the vehicle height could be approximated by measuring the relative displacement of the feature points within the group.

In this paper, we propose a vehicle shape approximation method from the vehicle motion over an image sequence. Our strategy is to collect more information, such as 3-D information, in the early stage in the expense of higher processing cost. In return, this extra information provides a more robust solution for the later stages such as vehicle tracking and classification in the real world conditions. In the next section, the overall concepts and methodology are summarized. Then, the four key stages including feature extraction, feature correspondence, height from motion, and vehicle reconstruction are described in details in Sections III to VI. Some results are depicted and discussed in Section VII.

0-7803-7194-1/01/$10.00 ©2001 IEEE

II. Methodology

In this paper, we assume the followings have already be computed from the image sequence:

1. Binary Vehicle Mask (*M*):

$$M(x,y)=\begin{cases}1, & \text{pixel (x, y) is vehicle pixel}\\ 0, & \text{otherwise}\end{cases}.$$

It is a binary representation of the vehicle extracted from the image sequence. The vehicle mask extraction is defined by subtracting the image sequence with the estimated stationary background [6].

2. Camera Parameters:

Camera parameters are important parameters that govern the relationship between the 2D image data and the 3D geometry data. With the required camera parameters, it is possible to reconstruct the vehicle. These parameters may include pan angle (*p*), tilt angle (*t*), swing angle (*s*), camera distance (*l*), and focal length (*f*). By observing the geometric properties of the road lane, an effective camera calibration technique which calibrate the required camera parameters relative to the road was proposed [7].

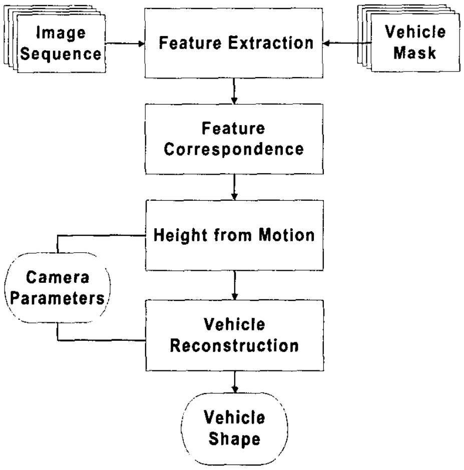


Figure 1. Proposed Method

The concept of the proposed method is to estimate the 3D shape of the vehicle from the vehicle motion over a traffic image sequence. Points with unique property that can be accurately tracking over image sequence are identified as feature points. By utilizing the property that the points on the same vehicle with different heights move at different speeds in the image coordinates, the heights can be estimated using the motion vectors of the feature points. Obviously, given the required camera parameters, the 3D shape of the vehicle can be approximated by the estimated 3D world coordinates of the feature points.

Basically, our proposed methodology is composed by four major stages as depicted in Figure 1, which are Feature Extraction, Feature Correspondence, Height from Motion and, Vehicle Reconstruction. Firstly, features with unique properties are extracted from the image sequence. It is important to identify the stable features over image sequence that can be confidently tracked in the next stage. Then, based on the features extracted from the last stage, the motion vectors of the features can be estimated by the given vehicle masks concerned. The motion vectors are then further refined by searching locally based on block matching. The motion vectors are then projected on the road plane based on the given camera parameters to compensate for the perspective effect. According to the ratios of the projected displacements of the features and the height of the camera, the height of each feature point relative to the road plane can be estimated. Finally, given the height of each features and the camera parameters, the mapping function from the 2D image coordinates to the 3D world coordinates is defined and the vehicle shape can be estimated.

III. Feature Extraction

In the feature extraction stage, the desired features with stable tracking over the image sequence should be identified. Features considered may include points, straight lines, curved lines and corners. Generally, corners are accepted as good features to solve feature point correspondence problem due to their uniqueness, simplicity and stability. In this paper, SUSAN corner detector [8] is used to detect corners of the vehicle. The basic idea of the SUSAN method is to associate to each pixel of the image a small area of neighbor pixels with similar brightness to this center pixel. From the size, centroid and axis of symmetry of these areas, corners or localized features are detected.

IV. Feature Correspondence

Block based motion estimation algorithm is adopted for this purpose. Basically, the displacement for a feature point (x_k, y_k) in the present frame k is determined by considering an $N_1 \times N_2$ block centered about (x_k, y_k) and searching frame $k+1$ for the best matching block of the same size. The search is usually limited by a maximum search distance s. In this paper, since the binary vehicle masks of frame k and $k+1$ are assumed to have been computed, the motion vector of the feature points can be initially estimated by the displacement (d_x, d_y) of the center of gravity of the mask. Then, the search area is now limited by a maximum search distance adding to this estimated motion vector, that is $(x_k + d_x \pm s, x_y + d_y \pm s)$. There are numerous criteria to quantify the matching of the blocks such as the minimum mean square error (MSE), the minimum mean absolute difference (MAD), and maximum matching pixel count (MPC). There are also many different searching strategies such as three-step search and cross search [9].

V. Height From Motion

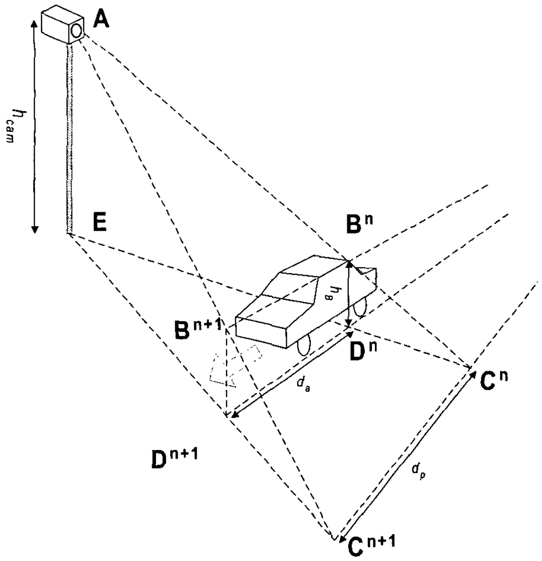


Figure 2. Geometry for Height Estimation from displacement between two image frames

By observation, points on the same vehicle but at different heights will move at slightly different speeds in the 2D image coordinates, where higher points move faster than lower ones. In Figure 2, the geometry for height estimation from feature point displacement between two image frames is depicted. In this figure, a vehicle is traveling in the direction of the arrow. A camera A is erected at the roadside point E and has a set of camera parameters that governs the position and viewing angles of the camera relative to the road.

For the symbols in the Figure 2, the superscript symbols denote the frame number, for example, B^n is the point B at frame n while B^{n+1} for frame $n+1$. At frame n, a feature point B of the vehicle is at the position B^n and is traveling in the arrow direction. At frame $n+1$, the point B is now at the position B^{n+1}. C^n and C^{n+1} are the projected position of B^n and B^{n+1} on the road plane respectively.

Let h_{cam} and h_B be the height of the camera and the height of point B. By similar triangles $\Delta C^n AE$ and $\Delta C^n B^n D^n$,

$$\frac{h_B}{h_{cam}} = \frac{B_n D_n}{AE} = \frac{C_n D_n}{C_n E}. \qquad (1)$$

Let d_a and d_p be the actual displacement and projected displacement of point B. By similar triangles $\Delta D^n D^{n+1} E$ and $\Delta C^n C^{n+1} E$,

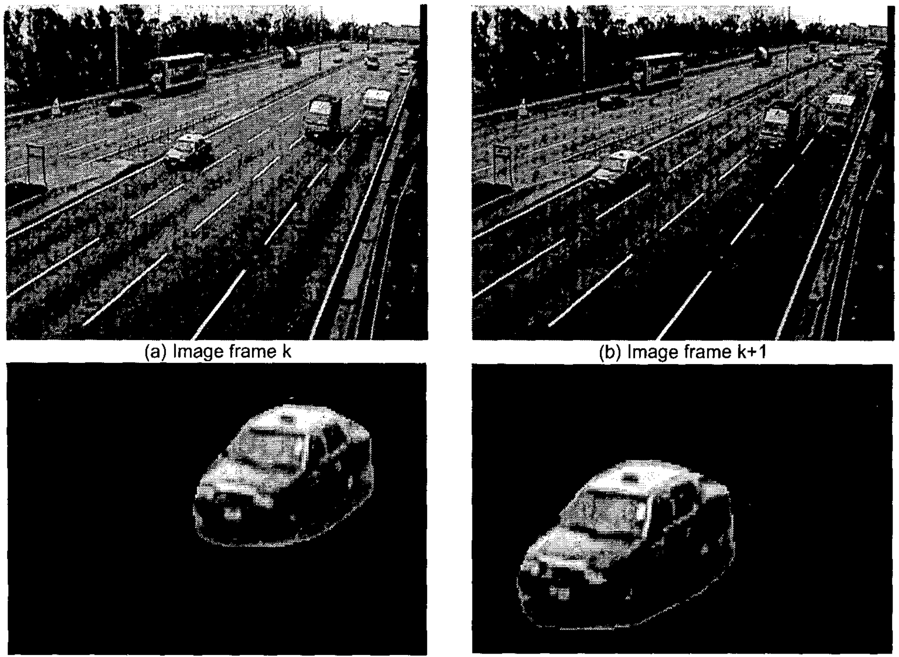
$$\frac{d_a}{d_p} = \frac{D^n D^{n+1}}{C^n C^{n+1}} = \frac{D^n E}{C^n E}$$

$$= \frac{C^n E^n - C^n D^n}{C^n E} = 1 - \frac{C^n D^n}{C^n E}. \qquad (2)$$

Solving (1) and (2),

$$h_B = h_{cam}(1 - \frac{d_a}{d_p}). \qquad (3)$$

To find h_B, the values of h_{cam}, d_p and d_a are required. h_{cam} is one of the camera parameters which is known. d_p can be computed by the mapping function defined in Appendix from the respective image coordinates and camera parameters. However, d_a can only be approximated by the shortest displacement measured for the feature group that belongs to the same vehicle. Basically, d_a corresponds to the feature point that is very close to the ground, such as wheel cover.



(a) Image frame k

(b) Image frame k+1

(c) Taxi mask at frame k (enlarged)

(d) Taxi mask at frame k+1 (enlarged)

Figure 3. Image sequence and vehicle mask at frame k and k+1

VI. Vehicle Reconstruction

In the last stage, the heights of all the feature points are computed. Given that image coordinates and the heights of the feature points, the world coordinates can be calculated by the inverse mapping function defined in the Appendix. To illustrate the reconstructed shape of the vehicle, one of the most effective methods is by showing the side view or the profile of the vehicle. Moreover, the dimensions, including length, width and height of the vehicle can also be approximated by the ranges of the world coordinates of the feature points in different axes. If there are sufficient feature points, more sophisticated methods, such as 3D surface fitting, can be used to illustrate the vehicle shape in more detail.

VII. Result Analysis

The image frame k and $k+1$ shown in Figure 3(a) and (b) are extracted from a 8-lane highway traffic video sequence captured at daytime. The frame rate of the video sequence is 5 frames/s. In Figure 3(c) and (d), the enlarged vehicle masks of frame k and $k+1$, where high brightness area representing the vehicle area, are shown *respectively.*

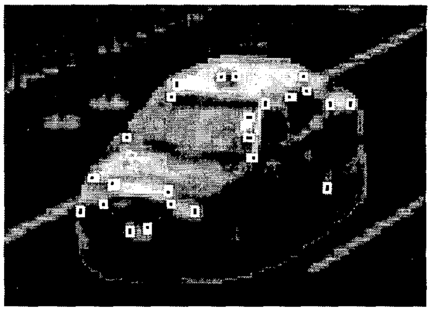


Figure 4. Corners detected by "SUSAN"

During feature extraction, "SUSAN" corner detector has detected 24 feature points for the taxi within the vehicle mask in frame k. These feature points, represented in the form of black dots surrounded by white squares, are depicted in Figure 4. By inspection, these feature points have high degree of uniqueness that are not easily confused with the neighboring pixels. Moreover, the feature points are distributed all over the taxi image and there are sufficient points at the upper part (top light) and the lower part (wheel, number plate) of the vehicles. There is no feature point at the homogeneous regions such as bonnet cover and vehicle top.

For feature correspondence, the motion vector of the center of gravity of the vehicle mask is shown in Figure 5(a). By using this motion vector, the motion vectors of the 24 features points are initially estimated and then further refined by searching locally using by a block-based matching method. Full search with MSE matching criteria is used. Since the motion vectors are accurately estimated, a small maximum search distance is used to significantly reduce the search area and computation cost. The block size and the search window size are typically set to 17×17 and 16×16 respectively. In Figure 5(b), the motion vectors of the 24 features are shown as 24 straight lines.

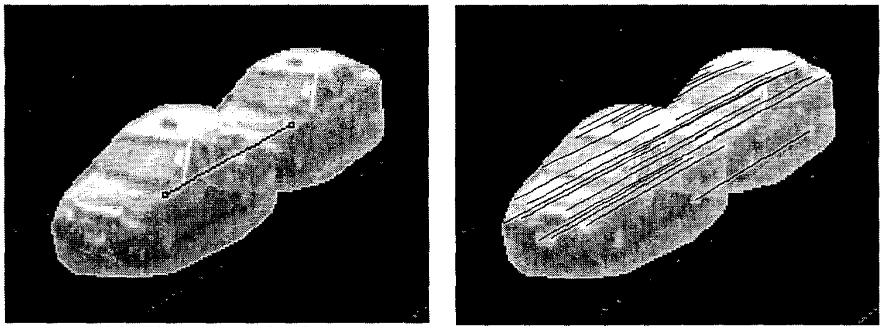
For height from motion, the 24 projected displacements computed from the motion vectors of the feature points are projected on to the road plane and depicted in Table 1. Among the 24 projected displacements, point 24, which is the feature point on the rear wheel, has the minimum projected displacement. It is used as the actual displacement reference. By equation (3), the heights are estimated from the projected displacements. Since point 24 is used as actual displacement reference, its height is 0m. The height of the number plate from the ground is around 0.3m to 0.5m. For the feature points on the number plate, point number 6 and 7, their heights are 0.3381m and 0.4079m respectively, which are in the suggested range.

Point	Projected Displacement (m)	Height
1	5.8054	0.2803
2	6.0830	0.7381
3	5.8932	0.4298
4	6.0302	0.6543
5	6.2124	0.9375
6	5.8391	0.3381
7	5.8802	0.4079
8	6.0995	0.7639
9	5.9457	0.5169
10	5.9906	0.5903
11	6.5166	1.3749
12	6.7262	1.6533
13	6.7042	1.6249
14	6.6794	1.5926
15	6.3048	1.0747
16	6.4407	1.2697
17	6.4271	1.2504
18	6.6472	1.5504
19	6.5929	1.4783
20	6.3695	1.1685
21	6.5633	1.4384
22	6.1924	0.9071
23	6.1761	0.8823
24	5.6476	0.0000

Table 1. Projected displacements and height of the figure points

	Length	Width	Height
Approximation	5.3640 m	1.8340 m	1.6533 m
Actual	4.8200 m	1.7650 m	1.6000 m
Error	11.29%	3.91%	3.33%

Table 2. Approximation of the taxi dimensions


(a) Initial motion estimation

(b) Motion vectors of feature points

Figure 5. Feature Correspondence

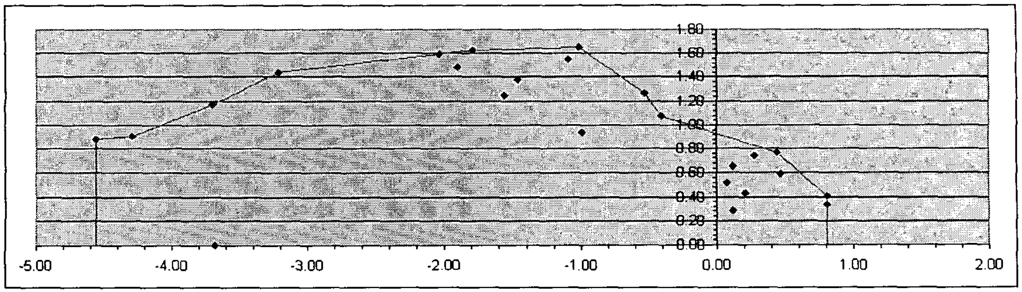


Figure 6. Taxi height profile or side view

For vehicle reconstruction, the world coordinates of each feature point are calculated from the image coordinates and the estimated height by the inverse mapping equations defined in Appendix. The height profile is depicted in Figure 6 by plotting the x and z coordinates of the feature points. The front of the taxi is facing right. To present a meaningful vehicle profile, straight lines are constructed from the feature point with maximum height to the next highest feature points and maintaining monotonic decrease along both directions. The bonnet, passenger and boot sections are clearly identified along the profile. Moreover, the length, width and height of the vehicle are estimated from the feature points are shown in Table 2 along with the actual dimension figures.

The source of the dimension estimation error may be caused by the spatial discretization of the image as the accuracy of the feature correspondence method can only up to pixel level. Moreover, since the height of the feature point is directly affected by the d_a/d_p, longer time interval between frames may reduce the effect of matching noise on the final results.

VIII. Conclusions

In conclusion, an estimation method of 3D vehicle shape information from a 2D monocular image sequence is presented in this paper. By extracting stable feature points, corresponding feature points between frames, computing the heights from projected displacements, and using the forward and backward image to world coordinates mapping function, the 3D vehicle shape can be sufficiently recovered. We have tested our method using real world traffic image sequence and the vehicle height profile and dimensions are estimated to be reasonably close to the actual dimensions. Currently, we are improving the accuracy of our method by using multiple frames or longer time interval and supersampling to minimize the spatial discretization effect.

IX. Appendix

The forward mapping, Φ, of a point, $\mathbf{Q} = (X_Q, Y_Q, Z_Q)$, in the world coordinates to a point, $\mathbf{q} = (x_q, y_q)$, from the image coordinates is defined as

$$\mathbf{q} = \Phi\{\mathbf{Q}\}, \tag{4}$$

where

$$x_q = \frac{f \cdot \begin{pmatrix} X_Q(\cos p \cos s + \sin p \sin t \sin s) \\ + Y_Q(\sin p \cos s - \cos p \sin t \sin s) \\ + Z_Q \cos t \sin s \end{pmatrix}}{-X_Q \sin p \cos t + Y_Q \cos p \cos t + Z_Q \sin t + l},$$

$$y_q = \frac{f \cdot \begin{pmatrix} X_Q(-\cos p \sin s + \sin p \sin t \cos s) \\ + Y_Q(-\sin p \sin s - \cos p \sin t \cos) \\ + Z_Q \cos t \cos s \end{pmatrix}}{-X_Q \sin p \cos t + Y_Q \cos p \cos t + Z_Q \sin t + l}. \tag{5}$$

The corresponding inverse mapping, Φ^{-1}, is defined as

$$\mathbf{Q} = \Phi^{-1}\{\mathbf{q}, Z_Q\}, \tag{6}$$

where

$$X_Q = \frac{\begin{pmatrix} \sin p(l + Z_Q \sin t)(x_q \sin s + y_q \cos s) \\ + \cos p(l \sin t + Z_Q)(x_q \cos s - y_q \sin s) \\ - Z_Q f \cos t \sin p \end{pmatrix}}{x_q \cos t \sin s + y_q \cos t \cos + f \sin t},$$

$$Y_Q = \frac{\begin{pmatrix} -\cos p(l + Z_Q \sin t)(x_q \sin s + y_q \cos s) \\ + \sin p(l \sin t + Z_Q)(x_q \cos s - y_q \sin s) \\ + Z_Q f \cos t \cos p \end{pmatrix}}{x_q \cos t \sin s + y_q \cos t \cos + f \sin t}. \tag{7}$$

and Z_Q is the assumed height of point **Q** in world coordinates. If point **Q** lies on the ground, Z_Q becomes zero.

X. References

[1] Hoose, N., *Computer Image Processing in Traffic Engineering*, Research Studies Press Ltd., London, 1991.

[2] Lai, H.S., "Vehicle Extraction and Modeling," *An Effective Methodology for Visual Traffic Surveillance*, Ph.D. Thesis, Chapter 5, The University of Hong Kong, 2000.

[3] Koller, D., Daniilidis, K., and, Nagel, H.H., "Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes," *International Journal of Computer Vision*, Vol. 10:3, pp.257-281, 1993.

[4] Huang, T.S., and Netravalli, A.N., "Motion and Structure from Feature Correspondences: A Review," *Proceedings of the IEEE*, vol. 82, no. 2, pp. 252- 268, February 1994.

[5] Malik, J., and, Russell, S., Traffic Surveillance and Detection Technology Development: New Traffic Sensor Technology Final Report, *California PATH Research Report UCB-ITS-PRR-97-6*, 1997.

[6] Fung, G.S.K., Yung, N.H.C., Pang, G.K.H., and Lai, A.H.S., "Effective Moving Cast Shadow Detection for Monocular Color Image Sequences,", *Research Report, ITS-2000G2*, Laboratory for ITS Research, The University of Hong Kong.

[7] Yung N.H.C., Pang G.K.H., Fung G.S.K., "A novel camera calibration technique for visual traffic surveillance", *Proc. 7th World Congress on Intelligence Transport Systems*, paper no.3024, 2000.

[8] Smith, S. M., and Brady, J. M., "SUSAN – A New Approach to Low Level Image Processing," *International Journal of Computer Vision*, 23(1), pp. 45-78, 1997.

[9] Tekalp, A.M., "Block-Based Methods", *Digital Video Processing*, Chapter 6, Prentice Hall, 1995.