Application of Planar Motion Segmentation for Scene Text Extraction

Tarak Gandhi, Rangachar Kasturi and Sameer Antani
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16801
{gandhi,kasturi,antani}@cse.psu.edu

Abstract

This paper explores an approach for extracting scene text from a sequence of images with relative motion between the camera and the scene. It is assumed that the scene text lies on planar surfaces, whereas the other features are likely to be at random depths or undergoing independent motion. The motion model parameters of these planar surfaces are estimated using gradient based methods, and multiple motion segmentation. The equations of the planar surfaces, as well as the camera motion parameters are extracted by combining the motion models of multiple planar surfaces. This approach is expected to improve the reliability and robustness of the estimates, which are used to perform perspective correction on the individual surfaces. Perspective correction can lead to improvement in OCR performance. This work could be useful for detecting road signs and billboards from a moving vehicle.

1 Introduction

There is a considerable amount of text occurring in video that is a useful source of information. The text that occurs naturally in the 3-D scene being imaged is called scene text. The scene text can have any orientation, and its image will be distorted by perspective projection in addition to being subject to the illumination conditions of the scene and susceptible to partial occlusion by other objects. There has been very little research on extracting scene text from general purpose video. The research that resembles this work the most is on recognition of vehicle license plates [4, 5]. However, these make restrictive assumptions about the text occuring in the scene.

Scene text typically exists on a planar surface in a 3-D scene. As the camera or the object moves, the motion of the text features should satisfy planar motion in 3-D. This research exploits this property to separate text features from features due to other objects which are likely to be at different random depths, and thus do not satisfy the planar constraint. A sequence of images can be used to segment different planar surfaces in the image, determine the model parameters, and remove outliers corresponding to clutter which do not fit any such surface, or is in motion with respect to these surface. The model parameters along with their estimated covariances can be used to determine the camera motion in terms of the linear and angular velocity, and the scene structure in terms of the plane normal equations. Since the camera motion parameters are the same for all planar surfaces, these parameters, as well as the plane normals of multiple planar surfaces are combined by using linear and non-linear methods. Using the estimated plane normals, the perspective effect of the camera on the characters can be compensated. This step would improve the accuracy of Optical Character Recognition (OCR).

2 Planar Motion Model

Let $X = (X_0, X_1, X_2)^t$ be the 3-D coordinates of a point in the camera coordinate system, in which the X_0 axis is the optical axis of the sensor. The perspective projection of the point in the image plane is given by:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{X_0} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \tag{1}$$

Let the relative motion between the camera and the scene be modeled by a translational velocity of $V = (V_0, V_1, V_2)^t$ and a rotational velocity of $W = (W_0, W_1, W_2)^t$. If the point lies on a planar surface with a normal along vector $K = (K_0, K_1, K_2)$ with an equation of $K^t X = 1$, the theoretical image motion such a point can be written as:

$$\begin{aligned} \hat{u}_1 &= a_1 + a_3 x_1 + a_5 x_2 + a_7 x_1^2 + a_8 x_1 x_2 \\ \hat{u}_2 &= a_2 + a_4 x_1 + a_6 x_2 + a_7 x_1 x_2 + a_8 x_2^2 \end{aligned} \tag{2}$$

Proceedings of the International Conference on Pattern Recognition (ICPR'00)
1051-4651/00 $10.00 @ 2000 IEEE

where the vector of eight coefficients $a = (a_1 \ldots a_8)^t$ called the planar motion model parameters can be expressed in terms of V, W, and K.

The 8-parameter equation given above is sufficient to represent the motion of a planar surface. However, the estimation of the quadratic parameters (a_7, a_8) may not be robust if small parts of the image are used. Hence, other models may be used in intermediate stages of segmentation. For example, the 6 parameter affine model – i.e., with $a_7 = a_8 = 0$ – is frequently used for motion segmentation.

3 Estimation of Model Parameters

The model parameters of planar motion can be estimated by using the image gradients and the optical flow constraint in parametric form as in [2]. Under favorable conditions, the temporal gradient g_0 and the spatial gradients g_1 and g_2 satisfy the following optical flow constraint equation:

$$g_0 + g_1 u_1 + g_2 u_2 = 0 \tag{3}$$

where $u = (u_1, u_2)^t$ is the image motion of the pixel. This constraint can be expressed as a minimization problem, and a least squares approach can be used to minimize the error in this constraint to estimate the motion model parameters, as well as their covariance Σ_a.

Direct application of this gradient based method yield accurate results only when the image motion of the pixels is less than 1 to 2 pixels per frame. To deal with larger image velocities, pyramid approach [2] is used. The model parameters are estimated from image gradients computed at the each resolution. These parameters are then used to warp the images at the next finer resolution. The process is repeated until the finest resolution is reached.

4 Multiple Motion Segmentation

Under ideal situations, where all image motion vectors result from a single planar surface motion and the noise in image gradients is Gaussian, a least squares fit approach gives an optimal estimate of the model parameters. However, even a few outliers, usually not belonging to the planar surfaces, can spoil the accuracy of the estimates. The problem is worse when the image contains multiple planar surfaces; the least squares approach if used directly, yields an 'average' estimate using all the surfaces. For obtaining reliable estimates of parameters, the image should be segmented into parts before applying the least squares. However, this is a chicken and egg problem, since the segmentation is what one wants to obtain by analyzing the image motion.

A considerable amount of work has been done on segmenting a scene into planar surfaces. Adiv [1] used Hough transform to map the image motion vectors into bins corresponding to the affine motion model. Bouthemy and Francois [3] use a two-term energy function, and a relaxation algorithm partitions the image into regions with different motion models. Black [2] uses a robust objective function with a multi-scale pyramid approach, and a spatial coherence constraint. In this ongoing work, the split and merge framework is being used for interactive segmentation.

5 Structure and Motion Parameters from Planar Surfaces

Using the motion model parameters (a), one can compute the structure (K) and motion (V, W) parameters using the well known method described in [8]. It should be noted that the internal parameters of the camera are required for determining the structure and motion parameters. There are two sets of linearly independent solutions for K and V, causing an ambiguity in structure and motion. In addition, there is an ambiguity due to the magnitude as well as the sign of the scale factor. The solutions corresponding a particular sign of the scale factor in each case correspond to objects lying behind the camera, and can be eliminated. The magnitude of the scale factor can then be overlooked by considering the solution where V is a unit vector. In the case of the scene containing multiple planar surfaces, the linear and angular velocities of the camera are identical for all the surfaces. If these parameters are determined by the above method, the constraint is not utilized, and each planar surface would give rise to different estimates of the camera motion. Furthermore, a single set of motion model parameters give rise to two sets of solutions for structure and motion parameters.

Use of image motion to directly estimate the structure and motion parameters have been previously proposed. Horn [6] describes the method to determine the camera motion as well as depth of all the scene points using the optical flow of the points for a general scene. Factorization methods [10] are also used to determine the structure and motion parameters from multiple frames, especially for orthographic and para-perspective motion models. However, it is noted that these methods require the knowledge of the full image motion. The direct use of image gradients instead of full optical flow makes the problem under-constrained [6], requiring smoothness constraints.

In this work, a piecewise planar model is assumed instead of a general scene to get the required smoothness constraint. The input to this algorithm are the motion model parameters of all the planar surfaces, computed independently using multiple motion segmentation. The algorithm combines these parameters to increase the accuracy of the structure and motion parameters, and removes the ambiguity due to multiple solutions.

1051-4651/00 $10.00 @ 2000 IEEE

Let the motion model parameters of each planar surface be given by a^i, and the respective plane normal vectors be K^i. These, as well as the camera motion parameters V and W can be stacked as follows:

$$\vec{A} = \begin{bmatrix} a^1 \\ a^2 \\ \vdots \\ a^n \end{bmatrix}, \vec{K} = \begin{bmatrix} K^1 \\ K^2 \\ \vdots \\ K^n \end{bmatrix}, \vec{L} = \begin{bmatrix} W \\ V \end{bmatrix}, \vec{P} = \begin{bmatrix} \vec{K} \\ \vec{L} \end{bmatrix} \tag{4}$$

The parameters a^i can be expressed in terms of V, W, and K^i as non-linear functions. However, these functions can be decoupled to obtain expressions in the following two forms for all planar surfaces $i = 1 \ldots n$:

$$\begin{bmatrix} G_2 & G_3(K^i) \end{bmatrix} \vec{L} = \vec{a}^i \tag{5}$$

$$G_1(\vec{L})\, K^i = \vec{a}^i - \begin{bmatrix} G_2 & O_{8\times 3} \end{bmatrix} \vec{L} \tag{6}$$

where G_1, G_2, and G_3 are 8×3 matrices, and $O_{8\times 3}$ is a zero matrix. If the dependencies in G's are neglected, the equations become linear in L and K^i, respectively, and can be iteratively solved using linear least squares. The system can be solved by solving equations 5 and 6 using linear least sqaures one after another repeatedly until convergence. The solutions obtained individually from planar surfaces can be used as starting solutions. However, the objective function which is optimized by this procedure does not correspond to the physical objective function to be minimized which is:

$$\left[\vec{A} - f(\vec{P})\right]^t \Sigma_A^{-1} \left[\vec{A} - f(\vec{P})\right] \tag{7}$$

subject to:

$$2\, c(\vec{P}) = V^t V - 1 = 0 \tag{8}$$

where f denotes the function which computes $\vec{A}$ from the structure and motion parameters combined in the vector $\vec{P}$. The above expression is minimized by iteratively incrementing $\vec{P_+}$, formed by stacking the Lagrange parameter λ to $\vec{P}$, using the following equation:

$$\Delta\vec{P}_+ = \begin{bmatrix} \Delta\vec{P} \\ \Delta\lambda \end{bmatrix} = \begin{bmatrix} Q & C^t \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} F^t\Sigma_A^{-1}\Delta\vec{A} \\ 0 \end{bmatrix} \tag{9}$$

where F and C denote the Jacobians of the functions f and c, respectively, and $Q = F^t\Sigma_A^{-1}F$. In actual practice, the non-linear iterations do not improve the accuracy significantly. However, the concept is useful for estimating the sensitivity of the parameters in $\vec{P}$. If the vector $\vec{A}$ has an error $\Delta\vec{A}$, the optimal solution changes by $\Delta\vec{P}$ to the first order approximation. From equation (9), using $E[\Delta\vec{A}\Delta\vec{A}^t] = \Sigma_A$, the covariance of $\vec{P}_+$ is approximately given by:

$$\begin{aligned} \Sigma_{P_+} &= E[\Delta\vec{P}_+\Delta\vec{P}_+^t] \\ &= \begin{bmatrix} Q & C^t \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q & C^t \\ C & 0 \end{bmatrix}^{-1} \end{aligned} \tag{10}$$

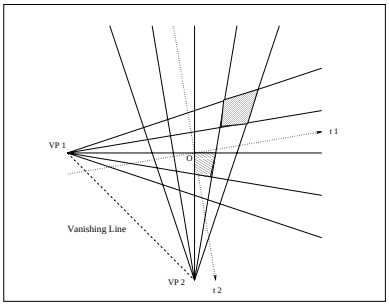

Figure 1. Distortion in perspective correction

The covariance of $\vec{P}$ is obtained by deleting the last row and column of Σ_{P_+} corresponding to the Lagrange parameter λ. For optimal estimation, the non-linear method can be applied after the linear method desscribed above. It should be noted that the velocity component V is normalized whenever it is updated, and the structure components K^i are correspondingly scaled.

Once the plane normal parameter for each surface is known, the surfaces can be rotated so as to face the camera axis. The rotation is performed around the axis perpendicular to the camera axis as well as the plane normal. This operation compensates for the perspective distortion. However, the rotation around the camera axis cannot be compensated using this method.

Furthermore, the effect of any error in the plane normal parameter results in larger error in certain parts of the image. This is illustrated in Figure 1, which shows the perspective distortion of a square grid due to the error in the estimate of the camera normal. The lines intersect at vanishing points whose distance from the origin is inversely proportional to the plane normal error. It can be seen that the distortion is worst at points that are far from the origin which corresponds to the point where the normal from the camera center intersects the plane.

6 Results

The application of the planar fit was tested on a number of simulated image sequences containing text or other patterns on planar surfaces. The translation, rotation and motion parameters of the camera, as well as the plane normal parameters were pre-specified.

The following experiment shows the result of estimation of plane normal, and its use in correcting the perspective distortion of the planar surfaces. The scene consists of two planar surfaces on both sides of the camera. The camera axis is parallel to these surfaces, and the camera is moving with a uniform linear velocity along its axis. Figure 2 (a) and (c) show the text surfaces on two sides of a camera, and an image frame obtained after projection is shown in Figure 2 (b). Using the sequence of images, the

1051-4651/00 $10.00 @ 2000 IEEE

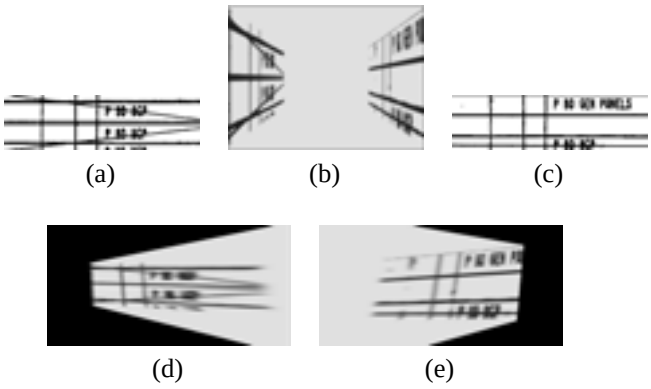

Figure 2. Perspective correction for simulated image sequence.

motion model parameters were obtained separately for the two halves of the image. The plane normal vectors were extracted from these parameters. The error in the plane normal estimate in terms of the angle between the estimated and actual normal vector is around $3.5°$ and $4.7°$ in the left and right planes respectively. The two parts of the image warped separately to correct for the perspective distortion. The corrected images are shown in Figure 2 (d) and (e). Although the error in the planar normal is small, the correction is somewhat imperfect, since the part of the plane that is imaged is very far from the normal through the camera center. However, it is much better than the perspective projection image of Figure 2 (b).

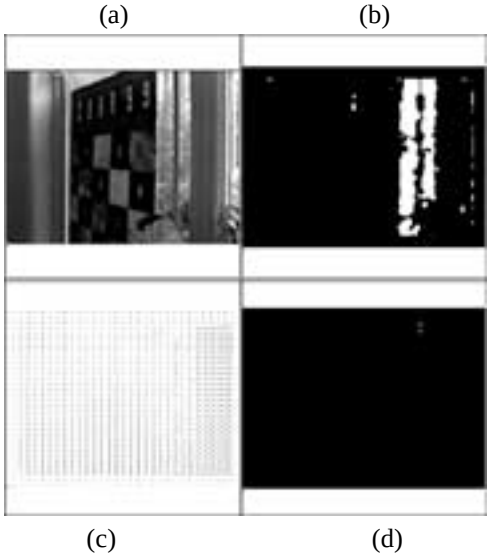

Figure 3. Segmentation of a real image sequence with poster containing text.

The segmentation procedure was tested on real image sequences captured using a hand-held video camera. A sample image in an indoor scene is shown in Figure 3 (a). The scene contains a poster with text characters, pasted on a window with the camera moving towards the poster. The background outside the window is at a large distance. In this scene, the segmentation picks up the poster as the dominant motion. In the next phase of segmentation, background is separated as another layer. However, background due to a perpendicular wall on the left is not separated, since it does not have enough features. The window bars are classified with the poster, since it is almost at the same depth as the poster. The segmentation is shown in Figure 3 (b). The image motion shown in Figure 3 (c) shows that the poster (except near the focus of expansion) as well as the window have larger motion than the distant background. Figure 3 (d) shows the detected outliers in each region which could make the parameter estimation unreliable. Figure 4 shows the corresponding results for an outdoor scene captured from a moving bus. The scene contains a "STOP" sign with a distant background. In this case, the segmentation picks up the background as the dominant motion, and separates the traffic sign in the next step. However, some pixels in the other parts of the image are mis-classified to belong to the planar surface, possibly because they may be lying close to the plane of the traffic sign. However, these pixels are isolated, and can be easily separated by performing connected component analysis and reclassifying small components. Also, the uniform areas of the sign are not classified properly, since the image motion in these regions cannot be estimated. Figure 5 (a) and (b) shows the images obtained after the perspective correction of the text segments in Figures 3 and 4, respectively. Focal length of the camera, given in the instruction manual was used for performing these corrections. In the case of the indoor scene, the correction is somewhat imperfect. The possible reasons could be the large distance of the plane normal through the camera from the field of view (see Figure 1), or an error in the focal length. The outdoor scene shows a satisfactory correction of the perspective distortion.

7 Summary and Future Work

This paper explores a novel approach to extract scene text from an image sequence with relative motion between the camera and the scene. The scene text was assumed to lie on planar surfaces, and planar motion model describing their motion was used to estimate the model parameters. The normal vectors of multiple planar surfaces with common camera motion were combined using linear and nonlinear approaches. These normal vectors were then used for correction of perspective distortion. This work can be useful in separating planar surfaces containing scene text from cluttered background, for example, to detect road signs and billboards from a moving vehicle.

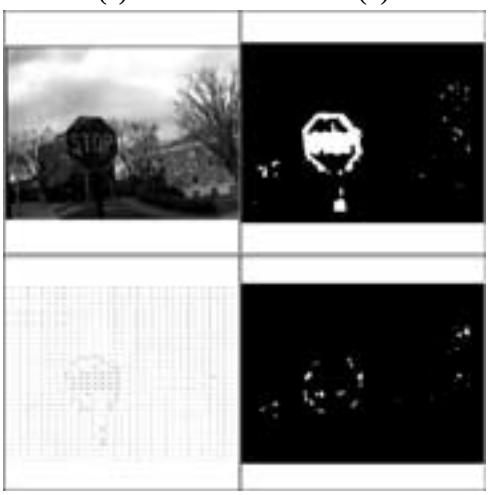

Figure 4. Segmentation of a real image sequence from an outdoor scene containing a "STOP" sign in distant background.

Figure 5. Resulting perspective correction

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 7(4):384–401, 1985.

[2] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.

[3] P. Bouthemy and E. Francois. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision*, 10(2):157–182, 1993.

[4] P. Comelli, P. Ferragina, M. N. Granieri, and F. Stabile. Optical recognition of motor vehicle license plates. *IEEE Trans. on Vehicular Technology*, 44(4):790–799, November 1995.

[5] Y. Cui and Q. Huang. Character extraction of license plates from video. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 502–507, 1997.

[6] B. K. P. Horn. *Robot Vision*. The MIT Electrical Engineering and Computer Science Series. The MIT Press, Cambridge, MA, 1986.

[7] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53(3):231–239, May 1991.

[8] K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier Science, Amsterdam, The Netherlands, 1996.

[9] S. Mann and R. W. Picard. Virtual bellows: Constructing high quality stills from video. In *Proc. IEEE International Conference on Image Processing*, pages I:363–367, November 1994.

[10] T. Morita and T. Kanade. A sequential factorization method for recovering shape and motion from image streams. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(8):858–867, August 1997.