VIDEO SYNTHESIS OF ARBITRARY VIEWS FOR APPROXIMATELY PLANAR SCENES

Amit K. Roy Chowdhury, Amit Kale and Rama Chellappa[*]

Center for Automation Research
University of Maryland
College Park MD 20742 USA
{amitrc,kale,rama}@cfar.umd.edu

ABSTRACT

In this paper, we propose a method to synthesize arbitrary views of a planar scene, given a monocular video sequence. The method is based on the availability of knowledge of the angle between the original and synthesized views. Such a method has many important applications, one of them being gait recognition. Gait recognition algorithms rely on the availability of an approximate side-view of the person. From a realistic viewpoint, such an assumption is impractical in surveillance applications and it is of interest to develop methods to synthesize a side view of the person, given an arbitrary view. For large distances from the camera, a planar approximation for the individual can be assumed. In this paper, we propose a perspective projection approach for recovering the direction of motion of the person purely from the video data, followed by synthesis of a new video sequence at a different angle. The algorithm works purely in the image and video domain, though 3D structure plays an implicit role in its theoretical justification. Examples of synthesized views using our method and performance evaluation are presented.

1. INTRODUCTION

In this paper, we propose a method to synthesize arbitrary views of approximately planar scenes, given a video sequence of the scene at any other viewing angle. The method we propose here has many applications in vision, video processing and multimedia. However, we were motivated from the point of view of gait recognition, which forms an important aspect of human identification [2, 3, 4, 5]. Thus, we will present the results of our method keeping in mind the gait recognition algorithm. The gait of a person is best reflected when he/she presents a side view (referred to in this paper as a canonical view) to the camera. Hence, most gait recognition algorithms rely on the availability of the side view of the subject. In realistic surveillance scenarios, however, it is unreasonable to assume that a subject would always present a side-view to the camera and hence, gait recognition algorithms need to work in a situation where the person walks at an arbitrary angle to the camera. The most general solution to this problem would involve estimating a 3D model of the person from which the required canonical view can be generated. This problem requires the solution of the structure from motion (SfM) or stereo reconstruction problems [6, 7], which are known to be notoriously hard. The situation is even more complicated when just a monocular video of the person is available.

Consider a person walking along a straight line which subtends an angle θ with the image plane (AC in Figure 1). If the distance, z_0, of the person from the camera is much larger than the width, Δz, of the person, then it is possible to replace the scaling factor $\frac{f}{z_0+\Delta z}$ for perspective projection by an average scaling factor $\frac{f}{z_0}$. In other words, for objects far enough from the camera, we can approximate the actual 3D object as being represented by a planar object. In this paper we show that, for planar or nearly planar scenes, it is possible to generate a canonical view given a monocular video sequence from any other view, in a way that uses the 3D structure only implicitly. Apart from the direct use of such canonical views for gait recognition, our method yields as a by-product, important information which can be useful its own right for different multimedia applications e.g. video compression, video indexing/retrieval etc.. We focus on the application of the theory to human gait in this paper.

In the context of gait recognition, several approaches have been proposed to circumvent the problems associated with the estimation of 3D model. Bobick and Johnson [8] had used linear regression to map static parameters across views. In [9], Shakhnarovich et al. computed an image based visual hull from a set of monocular views which was then used to render virtual views for tracking and recognition. This approach requires multiple calibrated cameras (at least 4) to synthesize the side-view. In contrast, our approach can work with only a single camera, involves a very simple calibration scheme and produces high quality synthesized videos. The computational complexity of the scheme is $O(mn)$, where $m \times n$ is the size of the bounding box around the person of interest.

We assume that we are given a video of a person walking at a fixed angle θ (Figure 1). We show that by tracking the direction of motion, α, in the video sequence, we can accurately estimate the angle θ in the 3D world. This can be done by using the perspective projection matrix. We also show that a simple, yet precise, camera calibration scheme can be designed for this problem. Under the assumption of planarity, using the angle θ and the calibration parameters, we can synthesize side-views or canonical views of the person, which can then be passed on to the gait recognition algorithms. Since the planar approximation is reasonable for many surveillance scenarios where the distance between the camera and people is large, this is a perfectly valid approach for synthesizing canonical views required by many gait recognition algorithms.

[*] Supported by the DARPA/ONR grant N00014-00-1-0908.

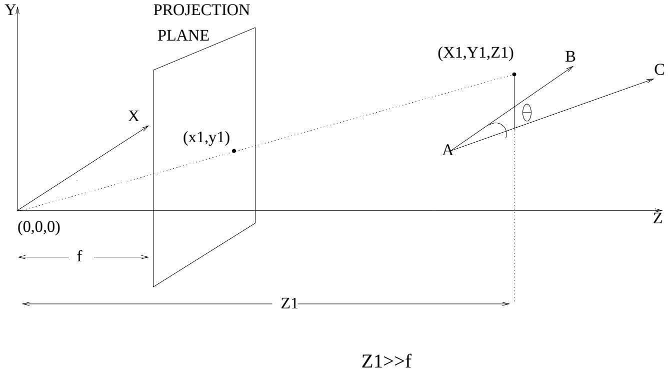

Fig. 1. Imaging Geometry

2. THEORY

2.1. Imaging Geometry

The imaging setup is shown in Figure 1(a). The coordinate frame is attached rigidly to a camera with the origin at the center of perspective projection and the z-axis perpendicular to the image plane. Assume that the person walks with a translational velocity $\mathbf{V} = [v_x, 0, v_z]^T$ along the line AC. The line AB is parallel to the image plane XY and this is the direction of the canonical view which needs to be synthesized. The angle between the straight line AB and AC, i.e. θ, represents a rotation about the vertical axis. Hence, we shall call this the azimuth angle. Our method works for $0 < \theta < 90$ degrees, though the performance deteriorates as θ increases. We will use the notation that $[X, Y, Z]$ denotes the coordinates of a point in 3D and $[x, y]$ its projection on the image plane.

2.2. Estimating the Azimuth Angle from Video Sequence

We assume that the person is walking along the straight line AC in Figure 1. Under exact perspective projection, straight lines map to straight lines. Thus the direction of motion in the 3D world corresponds to a straight line in the image plane, which can be estimated by tracking some points which move approximately rigidly as the person walks. Consider the equation of the 3D line which is at a height k from the ground plane and parallel to it, i.e.

$$Z = Tan(\theta)X + Z_0, Y = k. \tag{1}$$

Under perspective projection, this line transforms to

$$y = \frac{kf}{Z_0} - k\frac{Tan(\theta)}{f}x, \tag{2}$$

where $x = f\frac{X}{Z_0 + Tan(\theta)X}$, $y = f\frac{Y}{Z_0 + Tan(\theta)X}$ and f denotes the focal length of the camera. Thus if the slope of the line in the image plane, viz. $Tan(\alpha)$, is known, then given $K = -\frac{k}{f}$, the azimuth angle θ can be computed as

$$Tan(\theta) = \frac{1}{K}Tan(\alpha) \tag{3}$$

K can be obtained as a part of the calibration procedure. Note that using the orthographic projection model will result in giving a straight line $y = k$ which does not reflect the azimuth angle variation in the image plane. Thus our method will not work under orthographic projection assumptions.

2.3. Coordinate Transformation to Canonical View

Having obtained the angle θ, we need to synthesize the canonical view. Let Z_0 denote the distance of the object from the image plane (see Figure 1 (b)). If the dimensions of the object are small compared to Z_0, then $d\theta \approx 0$. This essentially corresponds to assuming a planar approximation of the object. Let $[X_\theta, Y_\theta, Z_\theta]'$ denote the coordinates of any point on the person who is walking at an angle θ to the image plane (as shown in the Figure 1(a)). Then

$$\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = R(\theta) \cdot \begin{bmatrix} X_\theta \\ Y_\theta \\ Z_\theta \end{bmatrix}, \tag{4}$$

where

$$R(\theta) = \begin{bmatrix} Cos(\theta) & 0 & Sin(\theta) \\ 0 & 1 & 0 \\ -Sin(\theta) & 0 & Cos(\theta) \end{bmatrix}. \tag{5}$$

Denoting the corresponding image plane coordinates as $[x_\theta, y_\theta]'$ and $[x_0, y_0]'$ (for $\theta = 0$) and using the perspective transformation, we can obtain the equations for $[x_0, y_0]'$ as

$$\begin{aligned} x_0 &= f\frac{x_\theta Cos(\theta) - fSin(\theta)}{fCos(\theta) + x_\theta Sin(\theta)} \\ y_0 &= f\frac{y_\theta}{-x_\theta Sin(\theta) + fCos(\theta)}, \end{aligned} \tag{6}$$

where

$$x = f\frac{X}{z} \text{and } y = f\frac{Y}{z}.$$

Equation (6) is particularly attractive since it does not involve the 3D depth; rather it is a direct transformation of the 2D image plane coordinates in the non-canonical view to get the image plane coordinates in the canonical one. Thus knowing the azimuth angle θ we can obtain a synthetic canonical view using (6) and a suitable texture mapping rule. Thus, for planar scenes, we are able to generate synthetic views purely from image data. This is important for many applications other than gait recognition e.g. multimedia.

The extension of the above method to synthesize arbitrary views is straight-forward. Suppose we are given a video sequence of a person walking at an angle θ_1. This can be estimated from the direction of motion of the person in the video sequence (as explained above). Once this is done we can synthesize the view at an angle θ_2 by applying the transformation of (6) with $\theta = \theta_2 - \theta_1$.

3. OBTAINING CAMERA CALIBRATION PARAMETERS

Use of (3) and (6) requires a knowledge of the parameters f and K, which are essentially the camera calibration parameters for this problem. In order to compute f, we used a calibration grid marked with 20 points. We placed the grid at 3 different azimuth angles $\theta = 15, 30, 45$ degrees and obtained the point correspondences by hand. The points are related by (6). We represent the three angles by $\theta_j \in \{15, 30, 45\}$ degrees and the coordinates of the i^{th} point by $[x^i_{\theta_j}, y^i_{\theta_j}]'$ and $[x^i_0, y^i_0]'$. Using these points we form a cost function $J(f)$ as shown in Equation (7). We solve this nonlinear regression using the Gauss-Newton method to obtain $f^* = \text{argmin}_f J(f)$.

$$J = \sum_{i,\theta_j} \left(x^i_0 - f\frac{x^i_{\theta_j}Cos(\theta_j) - fSin(\theta_j)}{fCos(\theta_j) + x^i_{\theta_j}Sin(\theta_j)} \right)^2$$

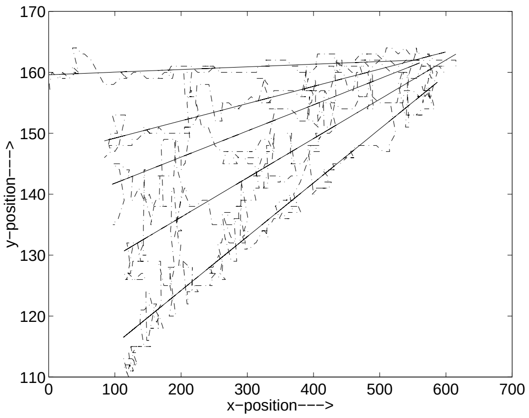

Fig. 2. Tracked points in the video sequence and the best fit straight lines.

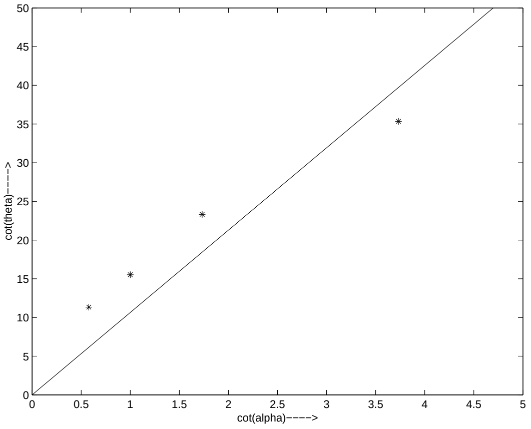

Fig. 3. Calibration curves for $Cot(\theta)$ vs. $Cot(\alpha)$. The dots represent the true values used for the calibration, while the straight line represents the best fit to these points using (3).

$$+\left(y_0^i - f\frac{y_{\theta_j}^i}{-x_{\theta_j}^i Sin(\theta_j) + fCos(\theta_j)}\right)^2. \qquad (7)$$

Next we consider the estimation of K. In order to do this, we captured videos of a person walking at $\theta = 0, 15, 30, 45, 60$ degrees. We tracked the position of a rigid point on the person followed by a median filtering of the trajectory. The resulting tracks are shown for the different θs in the Figure (2). To each of these tracks we fit a line using the least squares criterion. These are the solid lines in Figure (2), with slopes $Tan(\alpha(\theta))$. The top line is the case when $\theta = 0$. The lines for $\theta = 15$ is the one immediately below this line and so on. As may be expected, larger azimuth angles lead to larger image plane angles. The upper right corner where the lines intersect approximately, corresponds to the point from where the subjects start walking. These straight lines are the projections of the straight lines (one for each angle) traced out by the motion of the tracked rigid point in the 3D world. For the calibration procedure, we know the angle θ which traces out the straight line at the angle α. Given the corresponding values of α and θ, we can estimate K from (3). In Figure 3, we plot the values of $Cot(\theta)$ vs. $Cot(\alpha)$. The dots represent the true values used for the calibration, while the straight line represents the best fit to these points using (3).

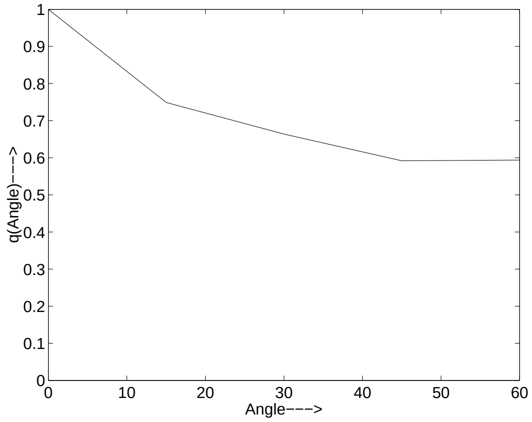

Fig. 4. Quality of Reconstruction as a function of the azimuth

4. EXPERIMENTAL RESULTS

In this section we describe our experimental setup and some of the synthesis results. People walk along straight lines at different values of azimuth angle $\theta = 0, 15, 30, 45, 60$ degrees. Background subtraction as discussed in [10] is first applied to the image sequence. To remove spurious noise, a standard 3×3 low-pass filter is applied to the resultant motion image. A bounding box is then placed around the part of the motion image that contains the moving person. The size of the box is chosen to accommodate the extreme cases of individuals in the database as regards height and girth. Subsequent processing is carried out inside this 'box'. For obtaining α we need to track a rigid point on the persons body. Since this is hard to accomplish, as an approximation we simply track the upper left corner of the box.

Given the video of an unknown person in the database, the above image processing operations are repeated to compute the image plane angle α. Using the calibration line shown in Figure 3, the azimuth angle θ was obtained. This is the most important step in the algorithm, since the estimation of θ is very sensitive to noise in the estimation of α. Using this value of θ and the value of f obtained as a part of the calibration procedure, the view of the person was synthesized using the Equations 6. Some of the syn-

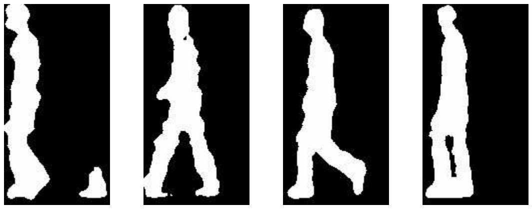

Fig. 5. Different stances of the person walking at azimuth angle $\theta = 0$

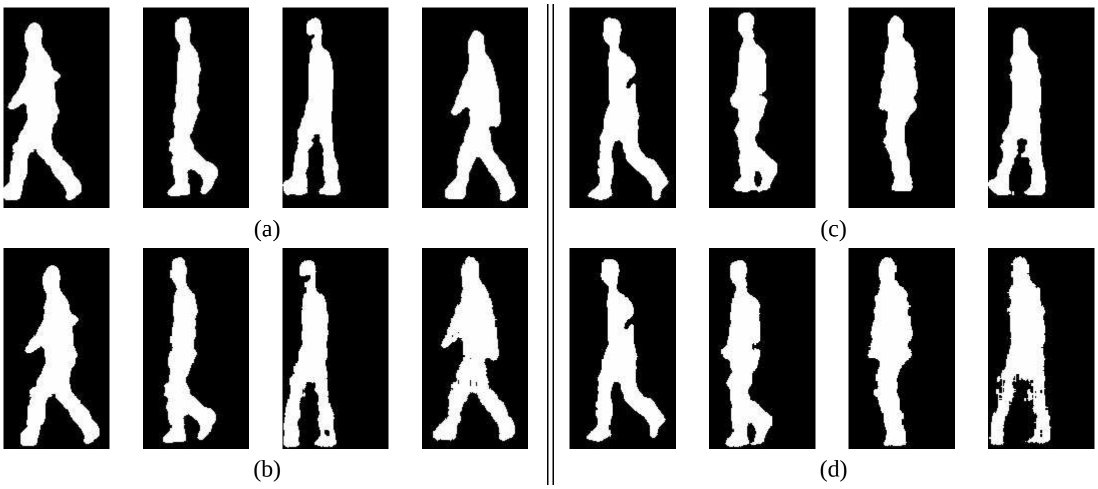

Fig. 6. (a) and (c) represent different stances of a person walking at angles 15 and 30 degrees to the camera; (b) and (d) represent side-views synthesized from original videos where the person walks at angles of 15 and 30 degrees to the camera.

thesis results are shown in Figure 6. It is difficult to represent the effect of the synthesis using just a few images. A few observations can be made, however. Note that in Figure 6 (a) and (c), the height of the person gradually decreases, indicating that he/she is moving away from the camera at a particular angle. On the contrary, the height of the synthesized silhouette is almost constant similar to the true zero azimuth case shown in Figure 5.

In order to assess the quality of the reconstructions we used the following idea. We take N contiguous boxed images of a person when he is walking at an azimuth $\theta = 0$. For every image transformed to the zero azimuth from the azimuth θ_x, we compute

$$q^{\theta_x} = \sum_{i=1}^{M} \text{maxcorr}(B_i^{\theta_x}, \mathcal{A}_N), \tag{8}$$

where B_i refers to the ith image in the synthesized sequence, $\mathcal{A}_N = \{A_1, \cdots, A_N\}$ is the set of N contiguous images for the zero azimuth, and

$$\text{maxcorr}(B_i^{\theta_x}, \mathcal{A}_N) = \max_j \frac{Num(A_j \bigcap B_i^{\theta_x})}{Num(A_j \bigcup B_i^{\theta_x})}.$$

(Num represents the number of overlapping pixels in the two images). The quality of the reconstruction, as a function of the azimuth angle is shown in the Figure 4. As expected, the performance degrades with increasing values of θ. The primary reason for this is the breakdown of the planarity assumption with increasing θ.

5. CONCLUSION

In this paper, we have proposed a method for synthesizing arbitrary views of planar objects. Our method uses a perspective projection model for estimating the azimuth angle of the original view from monocular video data. Thereafter, a video sequence at the new view is synthesized. The entire process is done in 2D, though 3D strucutre of the scene plays an implicit role. A simple, yet accurate, camera calibration procedure was also proposed. Examples of synthesized views are presented. Though the method has been explained from the motivation of the gait recognition problem, it has important aplications in other areas too, like multimedia and video processing. That forms a part of our future research into this problem.

6. REFERENCES

[1] J. Cutting and L. Kozlowski, "Recognizing friends by their walk:gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.

[2] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (Orlando, FL)*, May 2002.

[3] L. Lee and W.E.L. Grimson, "Gait analysis for recognition and classification," *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pp. 155–161, 2002.

[4] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognizing humans by gait via parametric canonical space," *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 359–366, October 1999.

[5] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," *Proceedings of IEEE Conference on Face and Gesture Recognition*, May 2002.

[6] O.D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[8] A.F. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[9] G.Shakhnarovich, L.Lee, and T.Darrell, "Integrated face and gait recognition from multiple views," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[10] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *FRAME-RATE Workshop, IEEE*, 1999.