VIDEO BASED RENDERING OF PLANAR DYNAMIC SCENES

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

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

ABSTRACT

In this paper, we propose a method to synthesize arbitrary views of a planar scene from a monocular video sequence of it. The 3-D direction of motion of the object is robustly estimated from the video sequence. Given this direction any other view of the object can be synthesized through a perspective projection approach, under assumptions of planarity. If the distance of the object from the camera is large, a planar approximation is reasonable even for non-planar scenes. Such a method has many important applications, one of them being gait recognition where a side view of the person is required. Our method can be used to synthesize the side-view of the person in case he/she does not present a side view to the camera. Since the planarity assumption is often an approximation, the effects of non-planarity can lead to inaccuracies in rendering and needs to be corrected for. Regions where this happens are examined and a simple technique based on weak perspective approximation is proposed to offset rendering inaccuracies. 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 dynamic 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 [1]. 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 even when 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 [2], which are known to be hard. Using the method described in this paper, we can synthesize any desired view without explicitly computing the 3-D structure. The method is based on the observation that if the distance, z_0, of the object from the camera is much larger than the width, Δz, of the object, 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. 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 in its own right for different multimedia applications e.g. video compression, video indexing/retrieval etc..

A comparison of our method to image-based rendering [3, 4, 5] is in order here. IBR approaches use images of a 3-D scene taken from appropriate locations of the camera to render an intermediate view. In this sense, usual image based rendering can be considered as an "interpolation". For the problem at hand, we seek to synthesize novel views given input images along a single viewing direction. In this sense the problem we are trying to handle can be regarded as an "extrapolation". For simplicity, we will be explaining with the canonical view (which happens to be the side-view) but it can be easily generalized to any other view as explained later.

We assume that we are given a video of a person walking at a fixed angle θ in the 3-D world(Figure 1). We show that by robustly tracking the direction of motion in the 2-D image sequence, α, we can accurately estimate the angle θ. 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 any other view e.g. the side-view or canonical views of the person, which can then be passed on to the gait recognition algorithms. One of the issues that needs to be considered is that when synthesized from a non-canonical view, certain parts, e.g. part of the persons back, not previously seen in the canonical view are visible. Ignoring the appearance of such unseen features during rendering can lead to inaccuracies. Regions where this happens are examined and a simple technique to offset rendering inaccuracies is proposed.

2. THEORY

2.1. Imaging Geometry

The imaging setup is shown in Figure 1. 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,

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

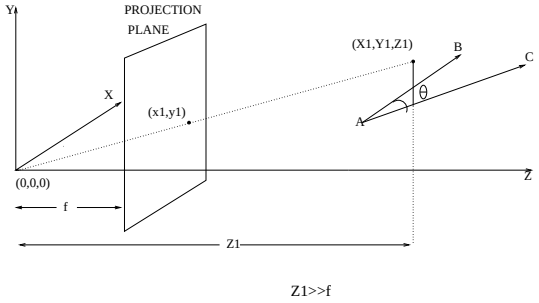

Fig. 1. Imaging Geometry

we shall call this the azimuth angle. 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

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

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

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. If the dimensions of the object are small compared to the distance from the image plane, it can be approximated as being planar, whereby a single value of azimuth θ can approximate the azimuth for every part of the body. 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). Then

$$\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = \begin{bmatrix} Cos(\theta) & 0 & Sin(\theta) \\ 0 & 1 & 0 \\ -Sin(\theta) & 0 & Cos(\theta) \end{bmatrix} \begin{bmatrix} X_\theta \\ Y_\theta \\ Z_\theta \end{bmatrix}, \tag{4}$$

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

$$\begin{aligned} x_0 &= f\frac{x_\theta Cos(\theta) - fSin(\theta)}{fCos(\theta) + x_\theta Sin(\theta)} \\ y_0 &= f\frac{y_\theta}{-x_\theta Sin(\theta) + fCos(\theta)}, \end{aligned} \tag{5}$$

where $x = f\frac{X}{z}$ and $y = f\frac{Y}{z}$. Equation (5) 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 (5) and a suitable texture mapping rule. Thus, for planar scenes, we are able to generate synthetic views purely from image data.

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 (5) with $\theta = \theta_2 - \theta_1$.

2.4. Compensation for non-planarity

When the imaged object is non-planar, portions unseen in the canonical view appear. An example is shown in Figure 2 which shows the top views of a rectangular block. The dashed rectangle represents the canonical case while the solid one represents the case when the block is at an angle θ. Under perspective projection, the length of projection of the tilted rectangular block on the image plane for $\theta \neq 0$ can be shown to be a function of the spatial coordinates of the block(see Appendix), while under a weak perspective projection it is a function of θ alone. To simplify analysis we used the weak perspective approximation. The error introduced by this approximation is small provided $Z >> a, b, x$. The dotted lines represent the true perspective projections of the edge of the block while the solid lines represent the weak perspective approximation. Under weak perspective projection, the apparent width of the block visible in the image plane is

$$W_a = \frac{f}{Z}(aCos(\theta) + bSin(\theta)). \tag{6}$$

Rendering this image of the block directly using Equation 5 will lead to an incorrect synthesized view. In general this problem can be avoided only if multiple cameras are available. For the present single camera case, a simple way to circumvent this problem is to synthesize only the portion of the image corresponding to $aCos(\theta)$.

3. OBTAINING CAMERA CALIBRATION PARAMETERS

Use of (3) and (5) requires a knowledge of the parameters f and K, which are essentially the camera calibration parameters for this problem. In order to compute f, a calibration grid marked with 20 points was placed at azimuth angles $\theta = 0, 15, 30, 45$ degrees. Point correspondences are obtained by hand and are related by (5).

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. Given the video of a person walking at an angle θ, we first need to estimate the image plane angle α. In order to do this we

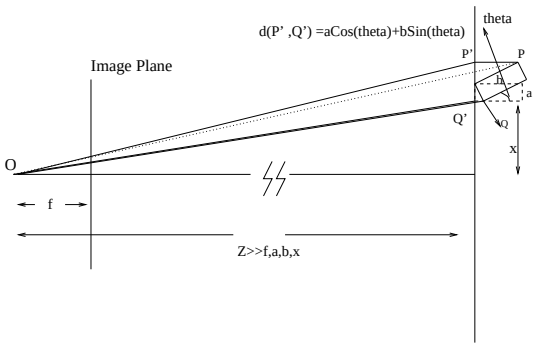

Fig. 2. Rendering of Non-planar objects

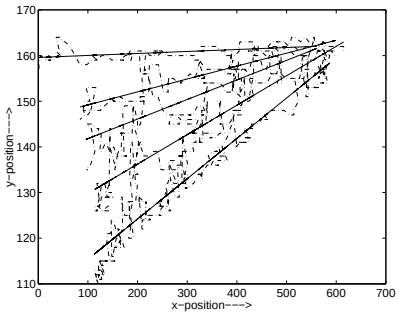

Fig. 3. Tracked points in the video sequence and the best fit straight lines.

place a bounding box around the person as explained in Section 4. This is approximately the same as tracking the position of a rigid point on the person. The resulting tracks are shown for the different θs in the Figure (3). Given point pairs $(x_i, y_i), i = 1 \cdots M$ we need to estimate the slope $A = Tan(\alpha)$ and the y-intercept B corresponding to a straight line passing through these points. A simple approach to do this would be to pick A and B to minimize the sum of squares of the residuals viz Minimize$_{A,B} \sum_{i=1}^{M} (y_i - Ax_i - B)^2$

We estimated A and B for each person in the database for each θ using least-squares. The result is shown in Figure 4(a). Clearly there is considerable overlap across (A, B) clusters. An error made in α would subsequently lead to an inaccurate estimation of θ via Equation 2. One of the reasons for the poor estimation is the presence of significant number of outliers in the tracks. A robust way to deal with this problem is to minimize the least median of squares, (LMEDS)viz.

$$\text{Minimize}_{A,B}\text{Median}_i(y_i - Ax_i - B)^2 \qquad (7)$$

The (A, B) clusters corresponding to the (LMEDS) estimation are shown in Figure 4 (b). The clusters in this case have greater separation compared to the least squares case. For one person the line thus estimated is shown as the solid lines in Figure (3), 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. 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).

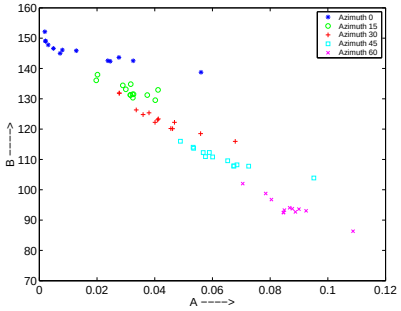

Fig. 4. (a) and (b) represents the clustering of A and B in Equation 7.

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 [6] is first applied to the image sequence. A bounding box is then placed around the part of the image that contains the moving person. As explained before the upper left corner of the box is tracked. Using LMEDS the image plane angle α was estimated. For an arbitrary walking path, a dynamic model needs to be used. Given K and α, the azimuth angle θ was obtained. Using this value of θ and f, the view of the person was synthesized using the Equations 5. A few results are shown in the Figure 7. It is observed that the torso appears wider than in the canonical view. As explained in Section 2.4 this is on account of the non-planarity of the torso region. The effects of non-planarity become more severe as the azimuth angle increases. To deal with this we need to ignore parts of the torso which are unseen in the canonical view. We assume the torso to be a rectangular block. Given the extent of the torso (U, L) the widths at different row positions $a(x) : U < x < L$ can be learned from different images of the person in the canonical view. Given θ, and an image in the non canonical view at azimuth θ only pixels upto a distance of $a(x)Cos(\theta)$ (see Equation 6) from the left extremity of the silhouette are synthesized using Equation 5. The next issue is how to obtain U and L. To this end, the width of the outer contour of the silhouette is computed for different images in the canonical view. Principal components analysis is performed on the width vectors, which are then projected on to the principal eigen vector. This has the effect of removing noise and retaining only the true variations which occur in the torso(hand)and leg regions. The result is shown in the Figure 5. From this overlay the extent of the torso region can be determined.

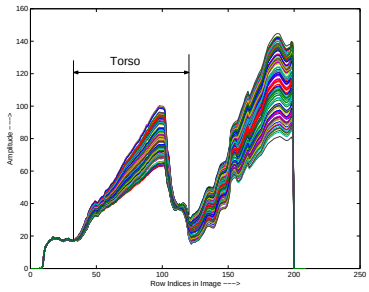

Fig. 5. Width Vectors projected on to the principal eigen vector for azimuth angle $\theta = 0$

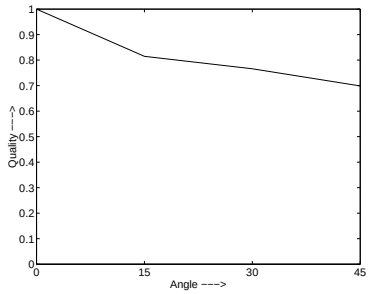

Fig. 6. Quality of reconstruction as a function of the angle of walk.

In order to assess the quality of the reconstructions we take the following approach. For every image synthesized from θ_x, we compute $q^{\theta_x} = \sum_{i=1}^{M} \text{maxcorr}(B_i^{\theta_x}, \mathcal{A}_N)$ where B_i refers to the ith image in the synthesized sequence, $\mathcal{A}_N$ is the set of N successive images in the canonical view and $maxcorr$ is the binary correlation [1]. The quality of the reconstruction, as a function of the azimuth angle is shown in the Figure 6. As expected, the performance degrades with increasing values of θ.

5. CONCLUSION

In this paper, we have proposed a method for synthesizing arbitrary views of planar objects moving along a straight line. Our method uses a perspective projection model and robust techniques to estimate the azimuth angle of the original view from monocular video data. Thereafter, a video sequence at the new view is synthesized. A simple technique to compensate for failures of non-planarity of the object is presented. The entire process is done in 2D, though 3D structure of the scene plays an implicit role. A simple, yet accurate, camera calibration procedure was also proposed. Examples of synthesized views were presented. One of future research tasks is to extend the method by using dynamic models for the direction of motion in the cases when the motion of the object follows arbitrary paths.

Appendix

Considering only the front face of the rectangular block in Figure 2 the projected length under true and weak perspective perspective projection are given by (8)and (9)as

$$W_{persp} = \frac{aCos(\theta) - \frac{xaSin(\theta)}{Z}}{1 + \frac{aSin(\theta)}{Z}} \quad (8)$$

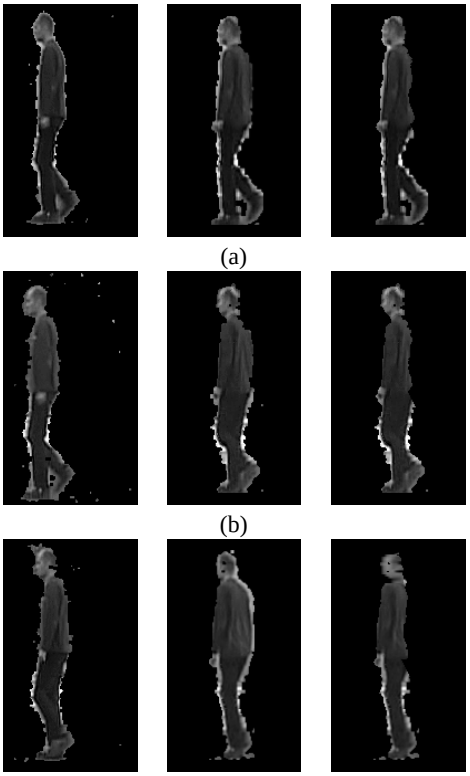

Fig. 7. The first column in (a) shows a frame from the canonical view; the second and third columns respectively show a image synthesized from a similar frame at an azimuth $\theta = 15$ before and after the correction for non-planarity is applied. (b) and (c) represent the same cases for azimuths $\theta = 30$ and 45 respectively.

$$W_{weakpersp} = aCos(\theta) \quad (9)$$

6. REFERENCES

[1] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," *Proc of the International Conference on Pattern Recognition*, 2002.

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

[3] S. Avidan and A. Shashua, "Novel view synthesis by cascading trilinear tensors," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 4, no. 4, 1998.

[4] Marc Levoy and Pat Hanrahan, "Light field rendering," *Computer Graphics*, vol. 30, no. Annual Conference Series, pp. 31–42, 1996.

[5] W. H. Leung and T. Chen, "Line-space representation and compression for image-based rendering," *Carnegie Mellon Technical Report*, vol. AMP01-02.

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