Synthesis of Novel Views of Moving Objects in Airborne Video

Zhanfeng Yue and Rama Chellappa
Center for Automation Research
University of Maryland
College Park, MD, USA, 20742
{zyue,rama}@cfar.umd.edu

Abstract

This paper presents a method for synthesizing novel views of moving objects in airborne video. The object of interest is tracked using an appearance based visual tracking method. The on-object point correspondence is then built and used to estimate the homography induced by the ground plane. With known camera focal length, the surface normal to the ground plane and the camera motion between two views are factored out from the homography. In order to assure robustness of surface normal estimation, a rank one constraint is applied to decompose a matrix which contains the homographies from multiple frame pairs. Given a desired viewing direction, the novel image of the object is generated by warping the reference frame using the new homography between the desired viewpoint and the reference frame. Experimental results show that the method is robust and errors due to small depth variations of the object is negligible.

1 Introduction

Object identification from surveillance video is of great importance for both military and civilian applications. The varying appearance of the object due to different viewing directions can make the identification very challenging, especially for template matching approaches. Generating novel views is an appealing solution to this problem, either one can generate templates for multiple viewing directions during the training stage, or the novel image of the probe object is rendered in the same viewing direction as some prebuilt templates during the testing stage. It can be accomplished by recovering the 3D information of the object from the available images and projecting to the desired view. Alternatively, image-based rendering techniques can be used, which rely on view interpolation or pixel reprojection and do not explicitly build a 3D model.

The problem of building 3D representations from a video sequence has been studied for more than twenty years [6, 8]. When frames from a single camera are used, one obtains a relative depth map from which novel views can be generated; estimates of absolute depth values can be obtained when multiple cameras are used. Reviews and comparisons of different Structure from Motion (SfM) methods that enable 3D modelling can be found in [2, 6, 8]. To build acceptable 3-D models, the following issues have to be considered:

BMVC 2005 doi:10.5244/C.19.36

observation noise (noise present in token correspondence or in computing optical flow), feature occlusion, motion/structure recovery ambiguities, mixed domain sequences having both small and large baselines and mismatched tokens and/or independently moving objects in the observed image frames. Being able to handle these issues is critical for producing practical structure recovery algorithms. Although elegant methods have been recently reported in [4, 13], much more needs to be done in addressing these issues. As an alternative to creating the 3-D model of an object prior to view synthesis, a novel view of the object for a desired viewpoint can be directly computed from a video sequence using the trilinear tensor [1], sidestepping the need for explicit 3D structure recovery. An accurate camera calibration is no longer necessary.

When the object to be synthesized is at a great distance from the camera (e.g., the vehicle in the airborne video), its depth-relief is negligible, and it may be assumed to move on a dominant plane (the ground plane). According to [14], if all the visible scene points are coplanar (i.e., structure degeneracy), the image data does not contain enough information to recover the epipolar geometry. Consequently, it is not possible to compute a 4×4 projective transformation between two sets of 3D points if the only correspondences available are coplanar. Therefore, the tensor based view synthesis method, which essentially needs the projective transformation between the 3D scene and the images, may not be accurate or stable. For the same reason, the ”Plane $+$ Parallax” approach [7] do not apply here since the object of interest is approximately a flat scene and the 3D structure can not be reliably estimated.

In order to accomplish the view synthesis task for moving objects in airborne video, we resort to homography induced by the ground plane using a synthesis by tracking algorithm. Our algorithm assumes as input a sequence of unstabilized images and the specified ROI. With the camera focal length available, the novel images of the object of interest are generated as the output according to the given viewing directions. In surveillance videos, tracking itself plays an important role in accumulating object information across frames. The region of interest (ROI, in our case, the moving object) is tracked with an appearance based visual tracking method. The tracking algorithm uses an adaptive appearance model incorporated in Sequential Monte Carlo (SMC) technique [16]. The on-object point correspondence is built from the tracking parameters and used to estimate the homography induced by the ground plane for each pair of frames. With known camera focal length, the surface normal to the ground plane and the camera motion between the frame pair are factored out from the homography using Singular Value Decomposition (SVD), as shown in [15]. We do not use some feature based robust approaches (e.g. the Harris corners and RANSAC [3]) to estimate the ground plane because of the low resolution and the lack of texture. With the tracking result across multiple frames and the estimated rotation between each frame pair available, a rank one constraint is applied to decompose a matrix that contains the homographies from multiple frame pairs [12] to ensure robust surface normal estimation. Given a desired viewing direction, the novel image of the object is generated by warping the reference frame using the new homography between the desired viewpoint and the reference frame.

The remainder of this paper is organized as follows: Section 2 briefly summarizes the appearance model-based visual tracking algorithm. Section 3 describes homography-based view synthesis for moving objects in airborne video, including a brief introduction to homography, establishing point correspondence from visual tracking, the estimation of ground plane surface normal and novel image generation for the desired viewing direction. In Section 4, we present a fusion scheme that utilizes the rank one constraint to realize robust surface normal estimation. Experimental results are given in Section 5, and Section 6 concludes the paper and discusses some future work.

2 Appearance Model-Based Visual Tracking

This section presents an appearance model-based tracking system for a single view. The system processes the video frames captured under one single view and produces the tracking parameters for later use. The task of an appearance tracker is to infer the deformation (or tracking) parameter best describing the differences between the observed appearances and the appearance model. To accommodate the dynamics embedded in the video sequence, we employ a state space time series model.

Suppose $\{Y_1, ..., Y_t, ...\}$ are the observed video frames containing the appearances of the object to be tracked. We use an affine transformation $\mathscr{T}$ parameterized by θ_t and denote the appearance model by A_t. Our time series model is fully defined by (a) *a state transition equation* and (b) *an observation equation.*

$$(a)\ \theta_t = \theta_{t-1} + U_t,\ (b)\ Z_t \doteq \mathscr{T}\{Y_t; \theta_t\} = A_t + V_t, \tag{1}$$

where U_t is the system noise and V_t is the observation noise. Our goal is to compute the posterior probability $p(\theta_t|Y_{1:t})$, which is used to estimate the 'best' parameter $\hat{\theta}_t$. Because this model is nonlinear (e.g. the affine transformation part), we use SMC technique [10, 11] to approximate $p(\theta_t|Y_{1:t})$ using a set of particles. We now specify the model choices.

2.1 Appearance Model A_t

The appearance model A_t is crucial in a tracker. If a fixed template, say $A_t \equiv A_0$, is used, it is difficult to handle appearance changes in the video. On the other hand, one could use a rapidly changing model, say $A_t = \hat{Z}_t \doteq \mathscr{T}\{Y_t; \hat{\theta}_t\}$, i.e., the 'best' patch of interest in the previous frame, but this is susceptible to drift. Thus, it is necessary to have a model which is a compromise between these two cases. Mixture models are used in [9, 16]. In this paper, we adapt the appearance model to the changing appearances at a moderate pace.

We assume that (i) the appearance model A_t is associated with a mean image μ_t (the actual A_t in (1)) and a variance image σ_t^2 (included in V_t in (1)), and (ii) A_t summarizes the past observations under an exponential envelop with a forgetting factor α. When the appearance in the current frame has been tracked, i.e. $\hat{Z}_t$ is ready, we compute an updated appearance model A_{t+1} and use it to track in the next frame. Using a maximum likelihood (ML) reasoning (skipped here due to space limitations), one can show that μ_{t+1} and σ_{t+1}^2 can be updated as

$$\mu_{t+1} = \alpha\mu_t + (1-\alpha)\hat{Z}_t;\ \sigma_{t+1}^2 = \alpha\sigma_t^2 + (1-\alpha)(\hat{Z}_t - \mu_t)^2. \tag{2}$$

In the above equations, all μ's and σ^2's are vectorized and the operation is element-wise. Also, V_t is distributed as a multivariate normal density $\mathscr{N}(0, D(\sigma_t^2))$, where $D(\sigma_t^2)$ denotes a diagonal matrix with diagonal elements σ_t^2.

2.2 Adaptive Noise U_t

The system noise U_t constrains the particle coverage. It is ideal to draw particles such that they are close to the object. In addition, the particle coverage should also accommodate the extent of clutter in the observation. To this end, we use $U_t \sim \mathcal{N}(\nu_t, r_t I)$, where ν_t is the 'instantaneous' velocity in the tracking parameter, r_t is the noise variance measuring the extent of clutter, and I is an identity matrix.

However, we have no knowledge of ν_t and r_t. We use a linear prediction scheme to estimate them. This prediction scheme is in spirit similar to finding an affine flows for the current 'best' patch in the next frame. Refer to [16] for details. As a consequence, the prediction scheme produces an estimate of ν_t and a prediction error ε_t. We take r_t as a monotone function of ε_t. Also, we vary the number of particles according to r_t.

3 Homography Based View Synthesis

3.1 Introduction to Homography

Suppose $\mathbf{X}$ is a scene point lying on a plane π. Let $\mathbf{x}$ and $\mathbf{x}'$ be the projections of $\mathbf{X}$ in view 1 and view 2 respectively. Then there exists a 3×3 matrix H_π such that $\mathbf{x}' \cong H_\pi \mathbf{x}$ where H_π is called the homography matrix induced by the plane π [5]. For simplicity we will omit the subscript of H_π if there is no confusion in the following sections.

Given a set of corresponding points $\mathbf{x_i} \leftrightarrow \mathbf{x_i'}$, where $\mathbf{x_i}$ come from view 1 and $\mathbf{x_i'}$ come from view 2, and writing $\mathbf{x_i'} = (x_i', y_i', \omega_i')^T$ with homogeneous coordinate, we can estimate the homography H between the two views using $\mathbf{x_i'} \times H\mathbf{x_i} = 0$ [5]. For each pair of corresponding points, three linear equations are written as

$$\begin{bmatrix} \mathbf{0}^T & -\omega_i' \mathbf{x_i}^T & -y_i' \mathbf{x_i}^T \\ \omega_i' \mathbf{x_i}^T & \mathbf{0}^T & -x_i' \mathbf{x_i}^T \\ y_i' \mathbf{x_i}^T & x_i' \mathbf{x_i}^T & \mathbf{0}^T \end{bmatrix} \begin{pmatrix} \mathbf{h_1} \\ \mathbf{h_2} \\ \mathbf{h_3} \end{pmatrix} = \mathbf{0} \tag{3}$$

where $\mathbf{h_i}, i = 1,2,3$ is a 3×1 vector made up of the entries in the i^{th} row of H.

By stacking the coordinates of all the corresponding points into a coefficient matrix A as in (3), the entries of H satisfy the linear equation $A\mathbf{h} = 0$ where $\mathbf{h} = (\mathbf{h_1}^T, \mathbf{h_2}^T, \mathbf{h_3}^T)^T$. More accurate result is expected using robust estimation methods (RANSAC [3], LMedS).

3.2 Building Point Correspondence from Visual Tracking

Since the height of the moving object in airborne video is negligible compared to the distance between the object and the camera, it may be assumed to move on the ground plane. Therefore the transformation which links the on-object points across frames can be approximated by a homography. To estimate the homography, the on-object point correspondence has to be found which is always challenging, especially for unstabilized video without rich texture. We use the appearance based visual tracking method introduced in Section 2 to establish the point correspondence. Without doing any stabilization, the moving object can be tracked even with the moving camera. The inter-frame transformation we obtain from the tracking algorithm captures the combined motion of the object and the camera. Without loss of generality, we pick the first frame as the reference frame wherein the ROI is specified by the user or by any detection algorithm. The ROI is tracked through the whole video sequence. Because of the rigidity of the object, all pixels inside the ROI undergo the same transformation so the point correspondence across frames can be found by sampling inside the ROI. (3) is then used to estimate the homography between the current frame and the reference frame.

3.3 Novel Image Generation for the Desired Viewing Direction

Assume that the projection matrix for the reference frame is $P_1 = K_1[I_{3\times3}|\mathbf{0}]$, where K_1 is the camera calibration matrix for the reference frame. For the specified viewing direction $[R_{new}|\mathbf{t_{new}}]$ relative to the reference frame, the homography H_{new} between the reference frame and the desired viewpoint induced by the ground plane is given by

$$H_{new} = K_{new}(R_{new} - \mathbf{t_{new}}\mathbf{n}^T)K_1{}^{-1}, \tag{4}$$

where K_{new} is the camera calibration matrix for the desired view, $\mathbf{n}^T$ is the surface normal to the ground plane in the coordinate system of the reference frame. Therefore, the on-object points $\mathbf{p}_i$'s in the reference frame and the corresponding points $\mathbf{p}_i'$'s in the desired view are related by $\mathbf{p}_i' = H_{new}\mathbf{p}_i$, which can be used to generate the desired image by warping the points from the reference frame.

3.4 Plane Function Estimation

In order to get H_{new} for a given $[R_{new}|\mathbf{t_{new}}]$ relative to the reference frame, the camera calibration matrices K_1 and K_{new}, and the ground plane surface normal $\mathbf{n}^T$ have to be known. By assuming that the principal point of the camera is at the center of the image and there is no skewing effect, the camera calibration matrix solely relies on the focal length f. As suggested in [5], f can be estimated using the inter-frame homographies H_k's and two imaged circular points $\mathbf{c}_j, j = 1,2$ in the reference frame if the calibration matrix is assumed to be constant ($K_k = K_1$) throughout the video sequence. In this paper, we simply obtain the focal length f from the metadata comes with the surveillance video.

It is not possible to get the ground plane information $\mathbf{n}^T$ from only one view. Triggs [15] gives an SVD based factorization method to decompose a calibrated homography $\hat{H} = K_2{}^{-1}HK_1$ into the plane normal $\mathbf{n}^T$ and the relative orientation between the two cameras $R(I_{3\times3}|-\mathbf{t})$. In the coordinate system of the first camera ($P_1 = (I_{3\times3}|0)$), let the 3D plane be $\mathbf{n}\times\mathbf{x} = z = 1/\zeta$, where $z = 1/\zeta > 0$ is the inverse distance to the plane. Let the matrix of the second camera be $P_2 = R(I_{3\times3}|-\mathbf{t})$ where $\mathbf{t}$ is the inter camera translation and R the inter camera rotation. Then the homography from image 1 to image 2 is $\hat{H} = R\hat{H}_1$ where $\hat{H}_1 = I_{3\times3} - \zeta\mathbf{t}\mathbf{n}^T$. For a 3D point $\mathbf{x}$ on the plane $\hat{H}\mathbf{x} = R(\mathbf{x} - \zeta\mathbf{t}\mathbf{n}^T) = R(\mathbf{x}-\mathbf{t}) \approx P_2\mathbf{x}$, since $\zeta\mathbf{n}^T\mathbf{x} = 1$ there. Treating $\mathbf{x}$ as a point in image 1 changes only the overall scale factor. Only the product $\zeta\mathbf{t}\mathbf{n}^T$ is recoverable, so we normalize to $||\mathbf{t}|| = ||\mathbf{n}|| = 1$ and use visibility tests to work out the allowable signs. The detailed decomposition of $\hat{H}$ can be found in appendix 1 of [15]. For a distant plane $\zeta \rightarrow 0$ as in airborne video, the estimated $\mathbf{n}^T$ and $\mathbf{t}$ might be unreliable but R is still accurate. The inaccuracy of $\mathbf{n}^T$ and $\mathbf{t}$ is compensated using multiple image pairs as described in Section 4, where the accurate R is used to compute the infinite homography H^∞.

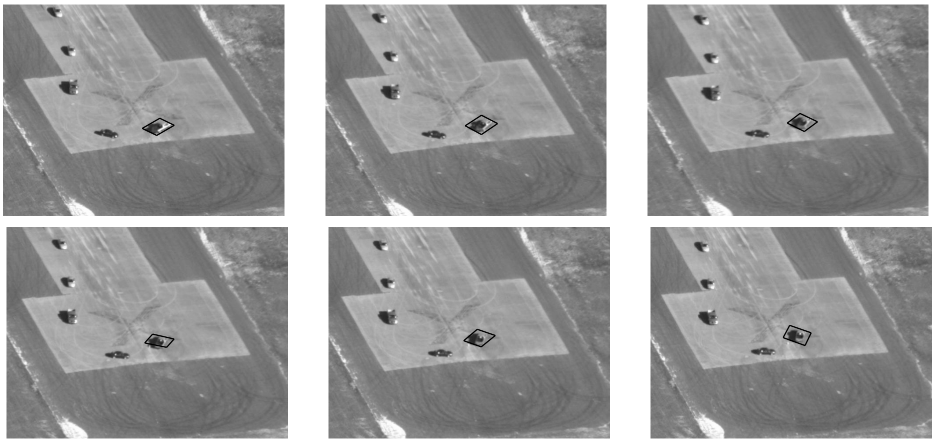

Figure 1: The appearance based visual tracking result, with the ROI marked as the black box in each frame.

4 Fusion Scheme

Every other frame, together with the reference frame, gives an estimate to the plane function $\mathbf{n}^T$ in the coordinate system of the reference frame. However, for a distant plane, the estimated $\mathbf{n}^T$ for each pair of frames is not reliable as pointed in [15]. Also, the information from the whole video sequence has not been fully utilized. Therefore, a suitable fusion scheme that can fuse available two-view estimates is needed to achieve a robust estimate of $\mathbf{n}^T$. In [12], a rank one constraint is applied to factorize a matrix, which stacks the planar homographies between the reference frame and all the other frames, into plane functions and camera motions. This is a good fit to our problem as a fusion scheme. The only information that is needed other than the inter-frame homography H_k is the infinite homography H_k^{∞} for each pair of frames. As mentioned in Section 3.4, for a distant plane the estimate to the camera rotation R_k is still accurate. Thus the infinite homography H_k^{∞} is computed as $H_k^{\infty} = K_k R_k K_1{}^{-1}$ [5]. A block matrix W is constructed by stacking all the transformed inter-frame homographies $\hat{H}_k$ as in (5). By applying the constraint that W has rank at most 1, W can be factorized into the camera center vector $[\bar{t}_k]$ and the ground plane surface normal $\mathbf{n}^T$ using SVD:

$$W = \begin{pmatrix} \hat{H}_2 \\ \hat{H}_3 \\ \vdots \\ \hat{H}_n \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{t}}_\mathbf{2} \\ \bar{\mathbf{t}}_\mathbf{3} \\ \vdots \\ \bar{\mathbf{t}}_\mathbf{n} \end{pmatrix} \mathbf{n}^T, \tag{5}$$

where $\hat{H}_k = \lambda^{-1} H_k^{\infty} H_k - I_{3\times 3}$. The scale λ for $\hat{H}_k$ is computed from the double eigenvalues of the planar homology $H_k^{\infty -1} H_k$.

Having the robust estimate of $\mathbf{n}^T$, we can use (4) to compute H_{new} and then the points on the object in the reference frame are warped to the desired viewpoint. A cubic interpolation is used to get the final synthesis result. With the metadata available, we can simply assume that the camera calibration matrix $K_{new} = K_k$ because mostly the desired image is in a comparable range of the available images, which relaxes the requirement that the focal length f be constant throughout the whole video sequence in Section 3.4.

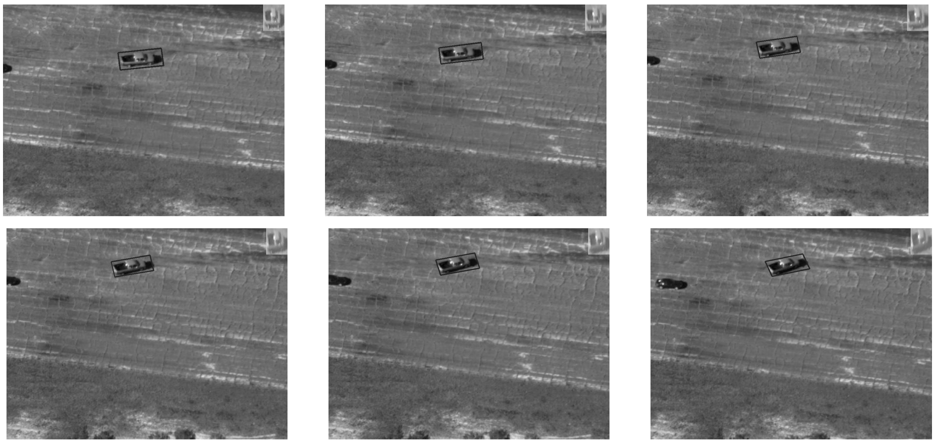

Figure 2: Another visual tracking result, with the ROI marked as the black box in each frame and the top right corner showing the appearance model updated at each frame.

The advantages of the proposed method include (i) Avoid the degeneracy in estimating the perspective projection relation across views. (ii) The desired viewpoint $[R_{new}|\mathbf{t_{new}}]$ is easy to be incorporated in the framework as shown in (4). (iii) The rank one constraint fusion scheme can help to improve the ground plane function estimation and view synthesis by using the information from the whole video sequence. (iv) No dense point correspondences are needed for view synthesis. (v) The computation is simple and fast.

5 Experimental Results

Experiments were conducted on several airborne video sequences. Fig. 1 shows the appearance based visual tracking result, with the ROI marked as the black box in each frame. We can observe from Fig. 1 that the object is reliably tracked in spite of its own motion and camera motion. Since the object is a rigid object, all pixels inside the ROI undergo the same transformation. Therefore the on-object point correspondences across frames are established by sampling inside the ROI and used to estimate the inter-frame homographies. Fig. 2 gives another example of the tracking result, with the top right corner showing the appearance model updated at each frame.

The view synthesis result are shown in Fig. 3 and Fig. 4, with the center image as the reference frame and the object inside the box. The surrounding images are the synthesized images w.r.t. different viewing directions. As we can observe, the synthesized images are very good in following the changing viewing directions although it is not easy to see the fine details because of the distance between the camera and the scene.

6 Conclusions and Future Work

A method for synthesizing novel views of moving objects in airborne video has been described in this paper. Correspondences among points on the object are built using an appearance based visual tracking method and used to estimate the homography induced by the ground plane. An SVD factorization is used to decompose the calibrated homography into the ground plane surface normal and the camera motion. A rank one constraint fusion scheme is used to realize robust surface normal estimation. New homography between a desired viewing direction and the reference frame is computed and used to generate novel images Experimental results show that the method is robust and the error due to the small depth variations of the object is negligible.

Our future work will use the proposed synthesis algorithm to build multiple templates during the learning stage, and identify objects using these templates to handle the various appearances due to different viewing directions. We will also do some experiments on simultaneous identification and tracking from video to video for airborne sequences.

References

[1] S. Avidan and A. Shashua, "Novel View Synthesis by Cascading Trilinear Tensors," *IEEE Transactions on Visualization and Computer Graphics*, 4(4):293–306, 1998.

[2] A. Chowdhury and R. Chellappa "Robust Estimation of Depth and Motion Using Stochastic Approximation," *Proc. ICIP*, 2001.

[3] M. Fischler and R. Bolles "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM*, Vol. 24, pp. 381–395, 1981

[4] D. Forsyth, S. Ioffe and J. Haddon "Bayesian Structure from Motion," *Proc. ICCV*, pp. 660–665, Corfu, Greece, 1999.

[5] R. Hartley, and A. Zisserman "Multiple View Geometry in Computer Vision ," Cambridge University Press, 2000.

[6] T. Huang and A. Netravali "Motion and Structure from Feature Correspondences: A Review," *Proc. IEEE*, 82:252–268, 1994.

[7] M. Irani, P. Anandan, and M. Cohen, "Direct Recovery of Planar-Parallax from Multiple Frames," *ICCV'99 Workshop: Vision Algorithms 99*, Corfu, Sept. 1999.

[8] T. Jebara, A. Azarbayejani and A. Pentland "3-D Structure from 2-D Motion," *IEEE Signal Processing Magazine*, 16: 66-84, 1999.

[9] A. Jepson, D. Fleet, and T. El-Maraghi "Robust Online Appearance Model for Visual Tracking," *Proc. CVPR,* 1:415-422,2001.

[10] G. Kitagawa "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models," *J. Computational and Graphical Statistics*, 5:1-25, 1996.

[11] J. S. Liu, and R. Chen "Sequential Monte Carlo for Dynamic Systems," *Journal of the American Statistical Association*, 93:1031-1041, 1998.

[12] C. Rother, S. Carlsson and D. Tell, "Projective Factorization of Planes and Cameras in Multiple Views," *Proc. ICPR*, Quebec, Canada, pp. 737-740, August 2002.

[13] S. Soatto and R. Brockett "Optimal Structure from Motion: Local Ambiguities and Global Estimation," *Proc. CVPR*, pp. 282–288, Santa Barbara, CA, 1998.

[14] P. Torr, A. Fitzgibbon and A. Zisserman "The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences," *International Journal of Computer Vision*, 32(1), 27–44, August, 1999.

[15] B. Triggs "Autocalibration from Planar Scenes," *Proc. 5th ECCV*, 1998.

[16] S. Zhou, R. Chellappa, and B. Moghaddam "Visual tracking and recognition using appearance-based modeling in particle filters," *IEEE Transactions on Image Processing*, 13(11), 1491-1506, Nov. 2004.

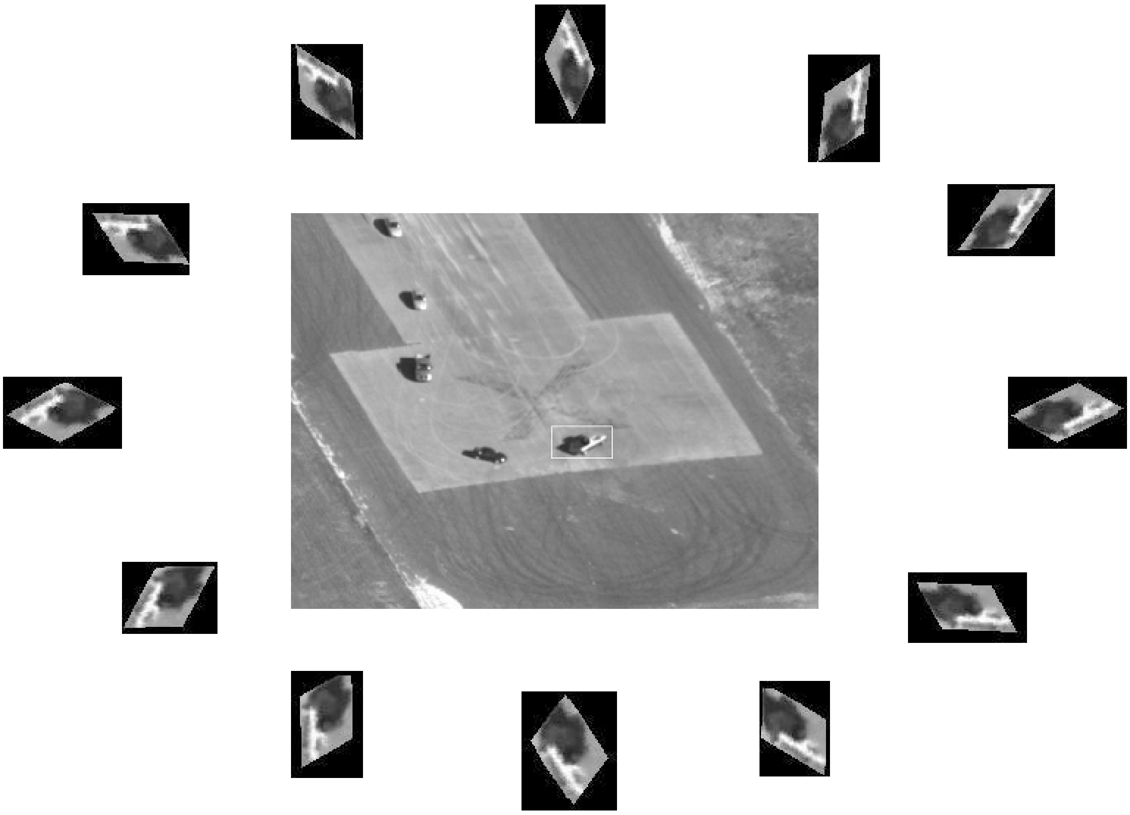

Figure 3: The view synthesis result. The center image is the reference frame, with the object inside the white box. The surrounding images are the synthesized images corresponding to different viewing directions.

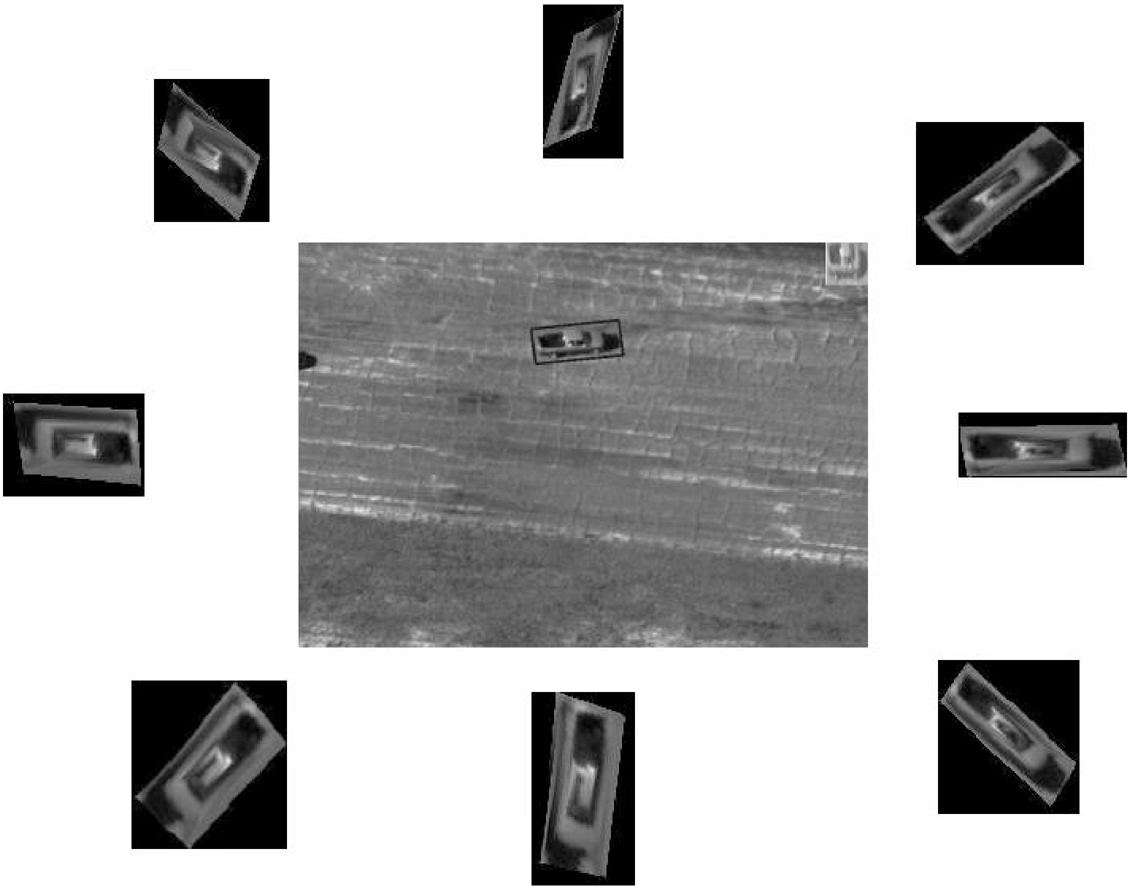

Figure 4: The view synthesis result. The center image is the reference frame, with the object inside the black box. The surrounding images are the synthesized images corresponding to different viewing directions.