Binocular Transfer Methods for Point-Feature Tracking of Image Sequences

Jason Z. Zhang, *Member, IEEE,* Q. M. Jonathan Wu, *Member, IEEE*, Hung-Tat Tsui, *Member, IEEE*, and William A. Gruver, *Member, IEEE*

***Abstract*—Image transfer is a method for projecting a 3D scene from two or more reference images. Typically, the correspondences of target points to be transferred and the reference points must be known over the reference images. We present two new transfer methods that eliminate the target point correspondence requirement. We show that five reference points matched across two reference images are sufficient to linearly resolve transfer under affine projection using two views instead of three views as needed by other techniques. Furthermore, given the correspondences of any four of the five reference points in any other view, we can transfer a target point to a third view from any one of the two original reference views. To improve the robustness of the affine projection method, we incorporate an orthographic camera model. A factorization method is applied to the reference points matched over two reference views. Experiments with real image sequences demonstrate the application of both methods for motion tracking.**

***Index Terms*—Affine camera model, feature tracking, image transfer, orthographic camera model.**

I. Introduction

FEATURE tracking is important in many applications of computer vision, including structure from motion [7], [8], [15], [17], [22], image synthesis [1], [11] and active motion tracking [4], [16], [17]. Image transfer is a method to perform feature tracking.

Image transfer has been developed for transferring scene points from known views. Previous research can be classified as geometric and algebraic. Geometric image transfer uses camera geometry relating the image measurements and the structure of the scene. The camera matrices and the shape of target objects usually must be recovered to perform image transfer. Mundy and Zisserman [10] established an image transfer model based on a linear representation using four control points with an affine camera model. Shapiro [17] designed an image transfer algorithm that was applied to clusters of reference points matched over sequential images. Tomasi and Kanade [22] applied a factorization method to calculate the intermediate camera matrices and the shape of the target object point to achieve affine image transfer. Reid, *et al.* [4], [16] applied a similar transfer method to active motion tracking.

With algebraic methods, recovery of camera matrices and object shapes is usually unnecessary. Instead, image transfer is performed with an "algebraic function of views" that involves only image coordinate measurements [18]. These algebraic functions, called *multiple view linearities* [2], [18], [23] and *multiple view tensors* [5], [6], [9], [19], [21], have linear relationships of the object features to multiple views and they are used to achieve direct image transfer within the image domain. Ullman and Basri [23] first showed that any three orthographic views of an object satisfy a linear function of the corresponding image coordinates. Shashua [18] extended the latter result to a more general form by showing that the linear function of three orthographic views is a particular case of a larger set of trilinear functions among three perspective views. The trilinear functions were represented by 27 coefficients that can be recovered using seven corresponding points over three images. When the coefficients are determined, the target point, whose corresponding coordinates between two reference images are known, can be directly transferred into the third image. Hartley [5] proposed a trifocal tensor method for transferring lines and points from two reference images into a third image. A $3 \times 3 \times 3$ trifocal tensor was shown to be identical to the coefficients of the trilinear functions introduced by Shashua. More recently, Kahl and Heyden [7] proposed *reduced third order centered affine tensors* to restrict the locations of corresponding points, lines, and conics across three views for shape and motion recovery. Their approach involved only 12 components in the reduced affine tensors. Shashua and Wolf [20] improved the three-view transfer method for points and lines using a *homography tensor* and its dual.

Transfer methods, regardless of their various forms, have required that correspondences of the target points be known in at least two reference views. This requirement is an obstacle that impedes the practical applications of transfer methods. For most computer vision applications, correspondence of the target points between images is difficult to ensure. In the case of shape from motion and "novel view" synthesis, for example, hundreds of points may have to be tracked over views, or an extended image sequence [1], [22]. In these situations, the target points may lie in a uniform intensity area of an image or they may be occluded by other objects in the views. Conventional feature detection methods can track a small number of distinct feature points. Thus, image transfer methods that rely on these algorithms are likely to fail.

In this research, we develop image transfer methods that eliminate the requirement of correspondence of the target points. Our

Manuscript received April 10, 2001; revised February 27, 2002 and September 27, 2002. This paper was recommended by Associate Editor I. Gu.

J. Z. Zhang and Q. M. J. Wu are with the Innovation Centre, National Research Council of Canada, Vancouver, BC V6T 1W5, Canada (e-mail: Jason.Zhang@nrc.ca, Jonathan.Wu@nrc.ca).

H.-T. Tsui is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China (e-mail: httsui@ee.cuhk.edu.hk).

W. A. Gruver is with the Intelligent Robotics and Manufacturing Systems Laboratory, School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: gruver@cs.sfu.ca).

Digital Object Identifier 10.1109/TSMCC.2002.806058

methods are developed under a framework of parallel projection that utilizes models based on affine camera projection and orthographic camera projection. As shown in the literature [7], [8], [10], [14]–[17], [22], [23], parallel camera projection models (affine and orthographic) are valid approximations to the perspective pin-hole camera model when the distance between the camera and an object is long enough so that the perspective effects in the imaging process become small. The simplicity of the parallel camera projection models is crucial for our methods.

We investigate a linear closed-form solution of image transfer under the affine camera projection model. Unlike conventional transfer methods, our proposed method can transfer a point from one image to another without having to find its correspondence in an intermediate image. Since the correspondence of the target point in the intermediate image is also determinable, our method can transfer all image points in the original image, including nondetectable image points, to any images in a given image sequence. This is valid even if the image of a target point is occluded, provided four control points for constructing a local affine coordinate frame (LCF) are known in the view. Because image transfer will be achieved between two rather than three views, we call it *binocular* transfer. We prove that four general points matched over three views, and one more point, whose image correspondence is known over two of the three views, are sufficient to achieve affine transfer. Because it needs only a small number of control points, our method is more suitable for dense point transfer than other techniques.

While simple and flexible, affine transfer is sensitive to noise in the image measurements and to changes in the configuration of the control points. Moreover, the affine camera model can introduce skew elements into the transferred images due to perspective effects. To alleviate the latter problem, we replace the affine camera model with an orthographic camera model. To reduce skew errors in the affine transfer, we introduce orthogonality into the orthographic camera. Noise sensitivity of the closed-form solution can be suppressed when more control points are available over three or more views, and a factorization method similar to that proposed by Tomasi and Kanada [22] is applied to the control points, resulting in a new transfer method based on a linear least-squares estimate. With the least-squares method, the 4-point LCF needed in the affine method is eliminated.

II. Affine Transfer: Binocular and Tri-Ocular Solutions

A. Points in an Affine Local Coordinate Frame

Given any four noncoplaner points $\boldsymbol{P}_i$ $(i = 0, 1, 2, 3)$ in $\boldsymbol{\mathcal{P}}^3$, the vectors

$$\boldsymbol{\varepsilon}_i = \boldsymbol{P}_i - \boldsymbol{P}_0,\ i = 1, 2, 3 \tag{1}$$

form a basis spanning $\boldsymbol{\mathcal{P}}^3$, where $\boldsymbol{P}_0$ is the origin of the basis. Therefore, any other point $\boldsymbol{P}_i \in \boldsymbol{\mathcal{P}}^3$, $i = 4, 5, \ldots$, can be linearly represented by

$$\boldsymbol{P}_i = \boldsymbol{P}_0 + \alpha_i\boldsymbol{\varepsilon}_1 + \beta_i\boldsymbol{\varepsilon}_2 + \gamma_i\boldsymbol{\varepsilon}_3 \tag{2}$$

where α_i, β_i and γ_i are the *affine coordinates* of point $\boldsymbol{P}_i$.

A 3D affine transformation is described by

$$\boldsymbol{P}'_i = \boldsymbol{A}\boldsymbol{P}_i + \boldsymbol{T} \tag{3}$$

where $\boldsymbol{A}$ is a 3×3 matrix and $\boldsymbol{T}$ a 3-vector.

Substituting (2) into (3) yields

$$\boldsymbol{P}'_i = \boldsymbol{P}'_0 + \alpha_i\boldsymbol{\varepsilon}'_1 + \beta_i\boldsymbol{\varepsilon}'_2 + \gamma_i\boldsymbol{\varepsilon}'_3 \tag{4}$$

where $\boldsymbol{P}'_0$, $\boldsymbol{\varepsilon}'_1$, $\boldsymbol{\varepsilon}'_2$, $\boldsymbol{\varepsilon}'_3$ are mappings of $\boldsymbol{P}_0$, $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$ under the affine transformation defined by (3), respectively.

As indicated by (2) and (4), the affine coordinates α_i, β_i and γ_i are geometric invariants under the 3D affine transformation. In general, the number of geometric invariants in a projection scenario equals the difference between the dimension of the geometric structure under viewing and the dimension of the transformation group acting on the structure [10], [13]. Since the 3D affine projection group is represented by a 4×4 matrix with the last row being $(0 \quad 0 \quad 0 \quad 1)$, it has $3 \times 4 = 12$ free parameters or degrees of freedom. Each point in $\boldsymbol{\mathcal{P}}^3$ has three degrees of freedom. Therefore, for a five-point structure in $\boldsymbol{\mathcal{P}}^3$, there are $3 \times 5 - 12 = 3$ absolute invariants α, β and γ. This suggests that affine invariants are potentially useful for visual tracking.

B. Affine Coordinates Computation: Two-View Solution

The affine camera model introduced by Mundy and Zisserman [10] is described by the 3D-2D transformation

$$\boldsymbol{p} = \boldsymbol{K}\boldsymbol{P} + \boldsymbol{t} \tag{5}$$

where $\boldsymbol{K}$ is a general 2×3 matrix representing the projection and orientation of a camera, $\boldsymbol{t}$ is a general 2-vector representing the displacement of the image between the two coordinate frames and $\boldsymbol{p}$ is the image of $\boldsymbol{P}$. By substituting (2) into (5) we obtain the linear representation of a point in the image plane

$$\boldsymbol{p}_i - \boldsymbol{p}_0 = \alpha_i\boldsymbol{e}_1 + \beta_i\boldsymbol{e}_2 + \gamma_i\boldsymbol{e}_3 \tag{6}$$

where $\boldsymbol{e}_j = \boldsymbol{K}\boldsymbol{\varepsilon}_j$, $j = 1, 2, 3$. A similar result holds for the transformed case

$$\boldsymbol{p}'_i - \boldsymbol{p}'_0 = \alpha_i\boldsymbol{e}'_1 + \beta_i\boldsymbol{e}'_2 + \gamma_i\boldsymbol{e}'_3. \tag{7}$$

The 3D affine transformation in (3) represents the motion of an object in space and the distortion in its shape. It is understandable that the effects of the object motion in the views of a stationary camera can be identically represented by the images taken by a moving camera viewing the same object as if it were motionless. In fact, for the transformed point $\boldsymbol{P}'$, we have from (5) that

$$\begin{aligned}\boldsymbol{p}' &= \boldsymbol{K}\boldsymbol{P}' + \boldsymbol{t} = \boldsymbol{K}(\boldsymbol{A}\boldsymbol{P} + \boldsymbol{T}) + \boldsymbol{t} \\ &= \boldsymbol{K}\boldsymbol{A}\boldsymbol{P} + (\boldsymbol{K}\boldsymbol{T} + \boldsymbol{t}) \\ &= \boldsymbol{K}'\boldsymbol{P} + \boldsymbol{t}'\end{aligned} \tag{8}$$

where $\boldsymbol{K}' = \boldsymbol{K}\boldsymbol{A}\boldsymbol{P}$ and $\boldsymbol{t}' = \boldsymbol{K}\boldsymbol{T} + \boldsymbol{t}$ represent the parameters of the camera at a new position. Therefore, without loss of generality, the identity between the motion of the object and the motion of the camera allows us to consider only one kind of motion. In the remainder of this paper, we shall consider only camera motion unless otherwise stated.

Given the affine coordinates α_i, β_i and γ_i of a point, its trajectory in an image sequence can be determined provided the affine LCF basis over the image sequence is known. We would like to compute the affine coordinates of a given point when the LCF is known. Observe from (6) and (7) that the affine coordinates of a point can be calculated from their corresponding image measurements across two views. Because the requirement of two-view correspondence could have disadvantages in certain applications, we want to compute the affine coordinates within a single reference view.

To calculate the three unknowns, without a two-view correspondence of the target point using (6) and (7), the affine epipolar relationship

$$ax' + by' + cx + dy + e = 0 \tag{9}$$

is employed, where a, b, c, d, and e are coefficients that denote the epipolar geometry of the two views under affine projection [17].

To obtain a non-trivial solution of (9), the correspondences of *four* non-coplanar points over the two images are needed, provided one of the five unknowns is a *free variable*. Without loss of generality, we assume e is a free variable, and set it to unity. The other unknowns a, b, c and d can be determined by resolving the system of linear equations using four matched points

$$ax'_i + by'_i + cx_i + dy_i = -1,\ i = 1, 2, 3, 4. \tag{10}$$

The constants a, b, c, d are the same for any matched points in the two views since they are only related to the camera parameters and the viewing positions of the camera. The condition that the four control points be non-coplanar ensures linear independence in (10).

Combining (6) and (7) with (10) yields the following system of equations in α_i, β_i, γ_i, x'_i and y'_i

$$\left.\begin{array}{c} ax'_i + by'_i + cx_i + dy_i = -1 \\ x_i = x_0 + \alpha_i e_{1x} + \beta_i e_{2x} + \gamma_i e_{3x} \\ y_i = y_0 + \alpha_i e_{1y} + \beta_i e_{2y} + \gamma_i e_{3y} \\ x'_i = x'_0 + \alpha_i e'_{1x} + \beta_i e'_{2x} + \gamma_i e'_{3x} \\ y'_i = y'_0 + \alpha_i e'_{1y} + \beta_i e'_{2y} + \gamma_i e'_{3y} \end{array}\right\} \tag{11}$$

where e_{ix}, e_{iy} and e'_{ix}, e'_{iy}, $i = 1, 2, 3$ are bases for the affine representations in (6) and (7).

Rewriting (11) in matrix form, we have

$$\boldsymbol{C}\boldsymbol{x} = \boldsymbol{b} \tag{12}$$

where $\boldsymbol{x} = (\alpha_i, \beta_i, \gamma_i, x'_i, y'_i)^\top$

$$\boldsymbol{C} = \begin{pmatrix} e_{1x} & e_{2x} & e_{3x} & 0 & 0 \\ e_{1y} & e_{2y} & e_{3y} & 0 & 0 \\ e'_{1x} & e'_{2x} & e'_{3x} & -1 & 0 \\ e'_{1y} & e'_{2y} & e'_{3y} & 0 & -1 \\ 0 & 0 & 0 & a & b \end{pmatrix}$$

$$\boldsymbol{b} = \begin{pmatrix} x_i - x_0 \\ y_i - y_0 \\ -x'_0 \\ -y'_0 \\ -1 - cx_i - dy_i \end{pmatrix}.$$

We have seen that four control points are needed to obtain the epipolar geometric parameters and the affine invariant coordinates. It is appealing to consider using the same set of control points to accomplish both computations. However, that would result in a reduction of the rank of $\boldsymbol{C}$ so that a unique solution of (12) could not be obtained. A condition ensuring that $\boldsymbol{C}$ is full-rank is given in the following result proved in the Appendix.

Lemma 1: $|\boldsymbol{C}| \neq 0$ if there exist at least five non-coplanar points.

Assuming there are at least five non-coplanar points, the unknown vector $\boldsymbol{x}$ in (12) can be represented by

$$\boldsymbol{x} = \boldsymbol{C}^{-1}\boldsymbol{b}. \tag{13}$$

From (13) we obtain (x'_i, y'_i), the image coordinates of the given target point in the second reference frame and $(\alpha_i, \beta_i, \gamma_i)$, the affine coordinates of the point at the same instant. This implies that image transfer can be achieved within two views. Consequently, a closed-form solution of the affine coordinates in the LCF can be obtained from Lemma 1 and (13) as follows:

Proposition 1: Five matched non-coplanar points across two views are sufficient to determine the affine coordinates of any other point visible in one of the two views.

C. Affine Transfer Among Two or More Views

Suppose there exists an image sequence $\{I_i | i = 1, 2, \ldots, M\}$ over which LCF's are tracked with four matched control points $\boldsymbol{p}_j^{(i)}$, $i \in \{1, 2, \ldots, M\}$, $j \in \{0, 1, 2, 3\}$. Let $\boldsymbol{p}_{1j}$ be an arbitrary point, other than the control points in the first image of the image sequence. If the affine coordinates of $\boldsymbol{p}_{1j}$, α_j, β_j, γ_j, have been obtained with the method proposed in Section II-B, then the counterpart of $\boldsymbol{p}_{1j}$ in any other view I_i, $i \in \{2, 3, \ldots, M\}$ can be determined as follows:

$$\boldsymbol{p}_{ij} = \boldsymbol{p}_0^{(i)} + \alpha_j \boldsymbol{e}1^{(i)} + \beta_j \boldsymbol{e}_2^{(i)} + \gamma_j \boldsymbol{e}_3^{(i)} \tag{14}$$

where $(\boldsymbol{e}_1^{(i)}, \boldsymbol{e}_2^{(i)}, \boldsymbol{e}_3^{(i)})$ is the affine LCF in the ith image.

Because the computation of the affine coordinates can be performed within any pair of images, for which the correspondences of the five control points are available, the original reference image is not necessarily the first frame in the image sequence. Instead, any frame in the sequence can be used as the reference from which a transfer process is computed. Therefore, feature points that are not visible in some, due to occlusion, but visible in other images can still be transferred throughout an image sequence. Furthermore, the transfer method is applicable for real-time motion tracking, whereby progressive updates of the local coordinate frames are made as the tracking process continues [16]. When any of the four control points forming the LCF's are not visible during the tracking process, additional control points can be obtained to construct new LCF's in the successive frames. The affine coordinates of target points must be recomputed with respect to the new LCF's. By this means, the image transfer of the previously tracked points can be continued.

III. Image Transfer Under Orthography

The affine transfer method, presented in Section II, provides an efficient means of binocular transfer. Since it is a least-control-point optimization solution, the method is control-point dependent, and particularly applicable to the cases where the conditions of point feature extraction are ideal, and the affine camera model (5) is valid. Furthermore, since the affine camera model is an un-calibrated model, skew effects may be caused in the transformed images by applying the affine model, particularly, in the case where perspective effects are present in the images [10], [17]. In this section, we address another binocular transfer method that employs a matrix factorization technique under orthographic projection. The method improves the robustness, and the image skew ability of the affine transfer method.

A. Sequential Image Representation Under Orthography

Let $\{(u_{fp}, v_{fp})|f = 1,2,\ldots,F,\ p = 1,2,\ldots,P\}$ denote the trajectories of P reference object points $\{\boldsymbol{s}_p = (x_p, y_p, z_p)^\top | p = 1,2,\ldots,P\}$ over F frames in an image sequence. Let the origin of the world-coordinate frame be defined as the centroid of the reference object points and let $(a_f,\ b_f)$ denote the centroid of the images of the reference points in frame f. Then it has been shown by Tomasi and Kanade [22] that a $2F \times P$ matrix $\tilde{\boldsymbol{W}}$

$$\tilde{\boldsymbol{W}} = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \cdots & \tilde{u}_{1P} \\ \tilde{v}_{11} & \tilde{v}_{12} & \cdots & \tilde{v}_{1P} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{u}_{F1} & \tilde{u}_{F2} & \cdots & \tilde{u}_{FP} \\ \tilde{v}_{F1} & \tilde{v}_{F2} & \cdots & \tilde{v}_{FP} \end{bmatrix} \tag{15}$$

where

$$\begin{aligned} \tilde{u}_{fp} &= u_{fp} - a_f \\ \tilde{v}_{fp} &= v_{fp} - b_f \end{aligned} \tag{16}$$

can be expressed as

$$\tilde{\boldsymbol{W}} = \boldsymbol{M}\boldsymbol{S} \tag{17}$$

where $\boldsymbol{M} = [\boldsymbol{i}_1 \quad \boldsymbol{j}_1 \quad \cdots \quad \boldsymbol{i}_F \quad \boldsymbol{j}_F]^\top$ encodes the coordinate axis orientations of the image planes of the orthographic camera, moving in space and $\boldsymbol{S} = [\boldsymbol{s}_1 \quad \boldsymbol{s}_2 \quad \cdots \quad \boldsymbol{s}_P]$ represents the structure of the object. The columns of $\boldsymbol{S}$ are the 3D coordinates of the reference points with respect to their centroids.

The registered image measurements $\tilde{u}_{fp}$ and $\tilde{v}_{fp}$ are related to the 3D coordinates of the point $\boldsymbol{s}_p$ by

$$\begin{bmatrix} \tilde{u}_{fp} \\ \tilde{v}_{fp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_f^\top \\ \boldsymbol{j}_f^\top \end{bmatrix} \boldsymbol{s}_p = \begin{bmatrix} u_{fp} \\ v_{fp} \end{bmatrix} + \begin{bmatrix} \boldsymbol{i}_f^\top \\ \boldsymbol{j}_f^\top \end{bmatrix} \boldsymbol{t}_f \tag{18}$$

where $\boldsymbol{t}_f$ is a vector from the world origin to the origin of the image frame f.

In the presence of image measurement noise, $\tilde{\boldsymbol{W}}$ can be approximated using singular value decomposition at the three largest singular values of $\tilde{\boldsymbol{W}}$ [3], [22]

$$\hat{\boldsymbol{W}} = \hat{\boldsymbol{M}}\hat{\boldsymbol{S}} \tag{19}$$

where $\hat{\boldsymbol{M}}$ is a $2F \times 3$ matrix and $\hat{\boldsymbol{S}}$ is a $3 \times P$ matrix.

Since the SVD of $\tilde{\boldsymbol{W}}$ is up to a nonsingular 3×3 matrix $\boldsymbol{A}$, the determination of $\hat{\boldsymbol{M}}$ needs an estimate of $\boldsymbol{A}$ such that

$$\begin{aligned} \boldsymbol{M} &= \hat{\boldsymbol{M}}\boldsymbol{A} \\ \boldsymbol{S} &= \boldsymbol{A}^{-1}\hat{\boldsymbol{S}}. \end{aligned} \tag{20}$$

Since $\boldsymbol{M}$ is a matrix denoting the actual rotation of the orthographic camera, and the rows of $\boldsymbol{M}$ are orthonormal, we have

$$\begin{aligned} \hat{\boldsymbol{i}}_f^\top \boldsymbol{A}\boldsymbol{A}^\top \hat{\boldsymbol{i}}_f &= 1 \\ \hat{\boldsymbol{j}}_f^\top \boldsymbol{A}\boldsymbol{A}^\top \hat{\boldsymbol{j}}_f &= 1 \\ \hat{\boldsymbol{i}}_f^\top \boldsymbol{A}\boldsymbol{A}^\top \hat{\boldsymbol{j}}_f &= 0. \end{aligned} \tag{21}$$

Define the symmetric matrix $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{A}^\top$. $\boldsymbol{Q}$ can be resolved by a linear-square method from the metric constraints on multiple views ($F \geq 2$) in (21). When it is positive-definite, $\boldsymbol{Q}$ can be parameterized by $\boldsymbol{G}\boldsymbol{G}^\top$, where $\boldsymbol{G}$ is a unique lower triangular matrix obtained from the Cholesky decomposition of $\boldsymbol{Q}$ [3]. Then $\boldsymbol{A}$ can be defined as $\boldsymbol{G}\boldsymbol{R}_1^\top$ (where $\boldsymbol{R}_1^\top$ is the rotation matrix of the camera at the first view) in the context of self-calibration [14].

In the presence of noise, the positive-definiteness of $\boldsymbol{Q}$ may be violated, and the Cholesky decomposition of $\boldsymbol{Q}$ may fail. Enumeration of different selections of the reference points to decompose the matrix in (19) from all available reference points can lead to a positive definite $\boldsymbol{Q}$ in many cases. An open problem, how to guarantee the positive-definiteness of $\boldsymbol{Q}$ in the enumeration method, still remains, however.

A more reliable method to tackle the Cholesky decomposition problem, is a non-linear minimization of the multi-view metric constraints in (21), imposed with the positive-definite constraint of $\boldsymbol{Q}$. To do so, assume $\boldsymbol{Q}$ is positive-definite, and is parameterized by $\boldsymbol{G}\boldsymbol{G}^\top$, where $\boldsymbol{G}$ is

$$\boldsymbol{G} = \begin{bmatrix} g_1 & 0 & 0 \\ g_2 & g_3 & 0 \\ g_4 & g_5 & g_6 \end{bmatrix}.$$

Because of the homogeneity of the non-singular matrix $\boldsymbol{A}$, the last diagonal element of $\boldsymbol{A}$ is assumed to be unity without loss of generality; therefore, we have $g_6 = 1$.

Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_5, 1)^\top$. The five independent parameters of the Cholesky parameterization of $\boldsymbol{Q}$ in $\boldsymbol{g}$ can be determined through the following minimization problem:

$$\begin{aligned} \min \phi(\boldsymbol{g}) = \sum_{i=1}^{F} \Bigg(& \left(\hat{\boldsymbol{i}}_f^\top \boldsymbol{Q} \hat{\boldsymbol{i}}_f - 1\right)^2 \\ & + \left(\hat{\boldsymbol{j}}_f^\top \boldsymbol{Q} \hat{\boldsymbol{j}}_f - 1\right)^2 + \left(\hat{\boldsymbol{i}}_f^\top \boldsymbol{Q} \hat{\boldsymbol{j}}_f\right)^2 \Bigg) \end{aligned} \tag{22}$$

where $\boldsymbol{Q} = \boldsymbol{G}\boldsymbol{G}^\top$.

An iterative non-linear least squares approach such as the Levenberg-Marquardt method [12] can be applied to solve the latter minimization problem.

The transformation $\boldsymbol{A}$ can be determined as

$$\boldsymbol{A} = \boldsymbol{G}\boldsymbol{R}^\top \tag{23}$$

where $\boldsymbol{R}_1^\top$ is the rotation matrix of the camera at the first view, through a simple self-calibration process presented in [14].

Knowing the camera matrix $\boldsymbol{M}$, we can represent the registered image measurements in frame f of a target point $\boldsymbol{s} = (x, y, z)^\top$ as

$$\tilde{u}_f = u_f - a'_f = \frac{P}{P+1}\boldsymbol{i}_f^\top \boldsymbol{s}$$

where P is the number of the reference points and

$$a'_f = \frac{\sum_{p=1}^{P} u_{fp} + u_f}{P+1}$$

is the centroid of all reference points plus the target point in the image. For $\tilde{v}_f$, we obtain a similar expression.

In summary, we obtain

$$\begin{bmatrix} \tilde{u}_f \\ \tilde{v}_f \end{bmatrix} = \frac{P}{P+1}\begin{bmatrix} \boldsymbol{i}_f^\top \\ \boldsymbol{j}_f^\top \end{bmatrix} \boldsymbol{s}. \tag{24}$$

Equation (24) is similar in form to (18) except the coefficient $P/(P+1)$ is introduced by adding the target point in the computation of the image centroid (a'_f, b'_f). When P is sufficiently large, $P/(P+1) \approx 1$ so that (18) can be substituted for (24).

B. Epipolar Constraint

The epipolar constraint in the affine projection method of Section II-B is now derived for the particular case of orthographic projection.

Rewrite the orthographic projection model, (18), as

$$\boldsymbol{x}_{fp} = \boldsymbol{M}_f \boldsymbol{s}_p + \boldsymbol{d}_f \tag{25}$$

where

$$\boldsymbol{x}_{fp} = \begin{bmatrix} u_{fp} \\ v_{fp} \end{bmatrix},\ \boldsymbol{M}_f = \begin{bmatrix} \boldsymbol{i}_f^\top \\ \boldsymbol{j}_f^\top \end{bmatrix},\ \boldsymbol{d}_f = -\boldsymbol{M}_f \boldsymbol{t}_f$$

By partitioning $\boldsymbol{M}_f$ as $(\boldsymbol{L}_f \mid \boldsymbol{l}_f)$ where $\boldsymbol{L}_f$ is a 2×2 matrix with orthonormal rows and $\boldsymbol{l}_f$ is a 2-vector, we express (25) as

$$\boldsymbol{x}_{fp} = \boldsymbol{L}_f \begin{bmatrix} x_p \\ y_p \end{bmatrix} + z_p \boldsymbol{l}_f + \boldsymbol{d}_f. \tag{26}$$

Similar partitioning performed on another view m leads to

$$\boldsymbol{x}_{mp} = \boldsymbol{L}_m \begin{bmatrix} x_p \\ y_p \end{bmatrix} + z_p \boldsymbol{l}_m + \boldsymbol{d}_m. \tag{27}$$

By eliminating the world coordinates $(x_p, y_p)^\top$, (26) and (27) yield a representation of the *epipolar line* between the two orthographic views

$$\boldsymbol{x}_{mp} = \boldsymbol{\Lambda}\boldsymbol{x}_{fp} + z_p \boldsymbol{\kappa} + \boldsymbol{\mu} \tag{28}$$

where $\boldsymbol{\Lambda} = \boldsymbol{L}_m \boldsymbol{L}_f^\top$, $\boldsymbol{\kappa} = \boldsymbol{l}_m - \boldsymbol{\Lambda}\boldsymbol{l}_f$ and $\boldsymbol{\mu} = \boldsymbol{d}_m - \boldsymbol{\Lambda}\boldsymbol{d}_f$. By definition, the quantities $\boldsymbol{\Lambda}$, $\boldsymbol{\kappa}$ and $\boldsymbol{\mu}$ depend only on the relative motion of the orthographic camera between the two views and are independent of the position of the target point.

Let $\boldsymbol{\kappa}^\perp$ be a vector perpendicular to $\boldsymbol{\kappa}$. Eliminating the depth z_p in (28) by using $\boldsymbol{\kappa}^\perp$ results in

$$(\boldsymbol{x}_{mp} - \boldsymbol{\Lambda}\boldsymbol{x}_{fp} - \boldsymbol{\mu}) \cdot \boldsymbol{\kappa}^\perp = 0 \tag{29}$$

which can be expressed in explicit form as

$$au_{mp} + bv_{mp} + cu_{fp} + dv_{fp} + e = 0 \tag{30}$$

where $(a, b)^\top = \boldsymbol{\kappa}^\perp$, $(c, d)^\top = -\boldsymbol{\Lambda}\boldsymbol{\kappa}^\perp$ and $e = \boldsymbol{\mu}^\top \boldsymbol{\kappa}^\perp$. Equation (30) has exactly a same form as the affine epipolar constraint equation in (9).

In the registered image coordinates, the homogeneous epipolar line (30) becomes

$$a\tilde{u}_{mp} + b\tilde{v}_{mp} + c\tilde{u}_{fp} + d\tilde{v}_{fp} = 0 \tag{31}$$

with four independent constants a, b, c, and d.

If we define $\boldsymbol{n} = (a, b, c, d)^\top$, determining the epipolar constraint is equivalent to resolving $\boldsymbol{n}$ using (30) or (31).

Although a non-trivial solution of $\boldsymbol{n}$ can be obtained from the image correspondences of a minimal set of four reference points, it has been shown in our experiments and those reported by Shapiro [17], that more points can generate a more robust and accurate solution. Therefore, we employ a linear least-squares estimate to obtain the epipolar vector $\boldsymbol{n}$.

Let $\tilde{\boldsymbol{x}}_{fp} = (\tilde{u}_{fp}, \tilde{v}_{fp})^\top$, $\tilde{\boldsymbol{x}}_{mp} = (\tilde{u}_{mp}, \tilde{v}_{mp})^\top$ and

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_{m1} & \tilde{\boldsymbol{x}}_{m2} & \cdots & \tilde{\boldsymbol{x}}_{mP} \\ \tilde{\boldsymbol{x}}_{f1} & \tilde{\boldsymbol{x}}_{f2} & \cdots & \tilde{\boldsymbol{x}}_{fP} \end{bmatrix}.$$

A scatter matrix is defined as

$$\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top. \tag{32}$$

Define a cost function for the least square estimate of $\boldsymbol{n}$ in (31) as

$$E(\boldsymbol{n}) = \frac{\sum_{p=1}^{P} (a\tilde{u}_{mp} + b\tilde{v}_{mp} + c\tilde{u}_{fp} + d\tilde{v}_{fp})^2}{a^2 + b^2 + c^2 + d^2} \tag{33}$$

which is related to the scatter matrix $\tilde{\boldsymbol{V}}$ and the epipolar constant vector $\boldsymbol{n}$ as:

$$E(\boldsymbol{n}) = \frac{\boldsymbol{n}^\top \tilde{\boldsymbol{V}} \boldsymbol{n}}{|\boldsymbol{n}|^2}. \tag{34}$$

Shapiro [17] showed that when $E(\boldsymbol{n})$ is minimum in a least-squares sense, $\boldsymbol{n}$ is the eigenvector corresponding to the minimum eigenvalue λ_1, satisfying

$$\tilde{\boldsymbol{V}}\boldsymbol{n} = \lambda_1 \boldsymbol{n}, \quad |\boldsymbol{n}|^2 = 1 \tag{35}$$

implying that

$$\min(E(\boldsymbol{n})) = \lambda_1.$$

With the above observations, we can obtain a minimum least-squares solution by solving for $\boldsymbol{n}$ in (35).

C. Least-Squares Method for Image Transfer

Given $\boldsymbol{x}_r = (u_r, v_r)^\top$, the image of a target point $\boldsymbol{s} = (x, y, z)^\top$ in a reference view r, we want to determine the trajectory of the target point $\boldsymbol{s}$ over the image sequence: $\{(u_f, v_f) \mid f = 1, 2, \ldots, F\}$.

To solve the transfer task, we propose the following steps.

1) Compute the camera matrix $\boldsymbol{M}$ in (20) using singular value decomposition of the registered measurement matrix $\tilde{\boldsymbol{W}}$ in Section III-A.
2) Calculate the 3D coordinates of the target point $\boldsymbol{s}$ to be transferred, whose image is $\boldsymbol{x}_r = (u_r, v_r)^\top$ in the reference view r.

3) Determine $\boldsymbol{x}_f = (u_f, v_f)^\top$, the image of the target point $\boldsymbol{s}$ in any view f, using (24) and (16).

Section III-A describes the methods required for the first step in the above image transfer. We present the solutions for Step 2 and Step 3 below.

Let a second reference view be m. Using the method presented in Section III-B, we compute the epipolar vector $\boldsymbol{n} = (a, b, c, d)^\top$ encoding the epipolar geometry between views r and m, which is expressed by (31).

We formulate the projections of $\boldsymbol{s}$ in the two views using the epipolar constraint. Equation (31) as follows:

$$\tilde{\boldsymbol{v}}_{mr} = \begin{bmatrix} \tilde{u}_m \\ \tilde{v}_m \\ \tilde{u}_r \\ \tilde{v}_r \end{bmatrix} = \rho \begin{bmatrix} \boldsymbol{i}_m^\top \\ \boldsymbol{j}_m^\top \\ \boldsymbol{i}_r^\top \\ \boldsymbol{j}_r^\top \end{bmatrix} \boldsymbol{s}$$

$$\tilde{\boldsymbol{v}}_{mr}^\top \boldsymbol{n} = 0 \tag{36}$$

where $\rho = P/(P+1)$, $\boldsymbol{i}_k^\top = (i_{k1}, i_{k2}, i_{k3})$ and $\boldsymbol{j}_k^\top = (j_{k1}, j_{k2}, j_{k3})$ $(k = r, m)$ are coordinate vectors of the image planes of the two views. The linear system in the five unknowns $\tilde{u}_m$, $\tilde{v}_m$, x, y, and z can be rewritten as

$$\boldsymbol{C}\boldsymbol{w} = \boldsymbol{b}. \tag{37}$$

where $\boldsymbol{w} = (\tilde{u}_m, \tilde{v}_m, x, y, z)^\top, \boldsymbol{b} = (0, 0, \tilde{u}_r, \tilde{v}_r, -c\tilde{u}_r - d\tilde{v}_r)^\top$ and

$$\boldsymbol{C} = \begin{bmatrix} -1 & 0 & \rho i_{m1} & \rho i_{m2} & \rho i_{m3} \\ 0 & -1 & \rho j_{m1} & \rho j_{m2} & \rho j_{m3} \\ 0 & 0 & \rho i_{r1} & \rho i_{r2} & \rho i_{r3} \\ 0 & 0 & \rho j_{r1} & \rho j_{r2} & \rho j_{r3} \\ a & b & 0 & 0 & 0 \end{bmatrix}$$

By simple manipulations, we obtain

$$|\boldsymbol{C}| = a\rho \begin{vmatrix} i_{m1} & i_{m2} & i_{m3} \\ i_{r1} & i_{r2} & i_{r3} \\ j_{r1} & j_{r2} & j_{r3} \end{vmatrix} + b\rho \begin{vmatrix} j_{m1} & j_{m2} & j_{m3} \\ i_{r1} & i_{r2} & i_{r3} \\ j_{r1} & j_{r2} & j_{r3} \end{vmatrix}. \tag{38}$$

Under the metric conditions in (21), $|\boldsymbol{C}| \neq 0$ unless the coordinate axes of the camera image planes become parallel due to camera motion. This is a rare case in practice. Therefore, we assume that $\boldsymbol{C}$ has full rank and is invertible. The 3D coordinates and the image in view m of the target point $\boldsymbol{s}$ are obtained by solving (37) as

$$\boldsymbol{w} = \begin{bmatrix} \tilde{u}_m \\ \tilde{v}_m \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{x}}_m \\ \boldsymbol{s} \end{bmatrix} = \boldsymbol{C}^{-1}\boldsymbol{b}. \tag{39}$$

Using the 3D coordinates of the target point, its image $\boldsymbol{x}_f = (u_f, v_f)^\top$ in any frame f can be obtained from (24) and (16):

$$\begin{bmatrix} u_f \\ v_f \end{bmatrix} = \frac{P}{P+1} \begin{bmatrix} \boldsymbol{i}_f^\top \\ \boldsymbol{j}_f^\top \end{bmatrix} \boldsymbol{s} + \begin{bmatrix} a_f \\ b_f \end{bmatrix} \tag{40}$$

where $f = 1, 2, \ldots, F$ and P is the number of reference points for the computation of the sequential image representation and the epipolar constraint.

IV. Experimental Results

We have applied the proposed transfer methods to track moving targets in real image sequences. The experiments described in this section, demonstrate the performance of the transfer methods for motion tracking.

A. Tracking by Affine Transfer

The first image sequence was captured using a CCD camera with a 25 mm lens mounted on an active platform, which allowed the camera to perform translational movements approximately 1.6 m from the target object—a model head. The size of the images in the sequence is 512×512 pixels. As shown below, the difference between the transferred coordinates and the real coordinates of target points is small. Therefore, the transfer method can be used to track objects represented by the transferred target points.

Fig. 1(a) shows a reference frame in the first sequence, from which target points are to be transferred over the image sequence. Twenty-two black dots on the face of the object were used as feature points. These points were numbered for identification.

To verify the accuracy of the affine transfer method, points 2, 11, 13, 21 were selected to compute the epipolar parameters, while points 2, 11, 13, 12 were used to construct the affine local coordinate frame (LCF). Table I lists the details of the transferred point coordinates, the measurements of the point coordinates and the errors of the transferred coordinates from the measurements in a third image. Similar results were obtained for the remaining frames of the image sequence. Table II lists the statistics for the transfer accuracy over the first three of the sequential images, where RMSV denotes the root-mean-square of variance of the absolute transfer errors. About 90% of the transferred points have sub-pixel accuracy and the average transfer error is also less than one pixel.

There are two sources of the transfer error. One source is measurement error in the feature point detection. The measurement error can be assumed to be Gaussian since no factor dominated the measurement accuracy. Another error source comes from the assumptions of the affine camera model, which only models the imaging processes well when the object extent in depth is small, compared to the distance between the object and the camera. When the conditions for the affine camera model assumption are violated, systemic error will result. This point will be verified in the orthographic transfer experiment.

Any improvement in feature point detection accuracy would reduce the transfer error of the points. The second type of error can be reduced either by ensuring the camera model assumptions, or by employing camera models such as weak-perspective or para-perspective [17] that accommodate the perspective effects in the imaging processes.

To assess the transfer performance of the method, the edges obtained with a LOG (Laplacian of Gaussian) edge detector in the reference image were tracked over the image sequence using the affine transfer method. Fig. 2 illustrates the results of the edge image transfer. The edge points detected in Fig. 2(a) were transferred into the sequential images shown in Fig. 2(b), (c) and (d). The transferred edge images were superimposed in

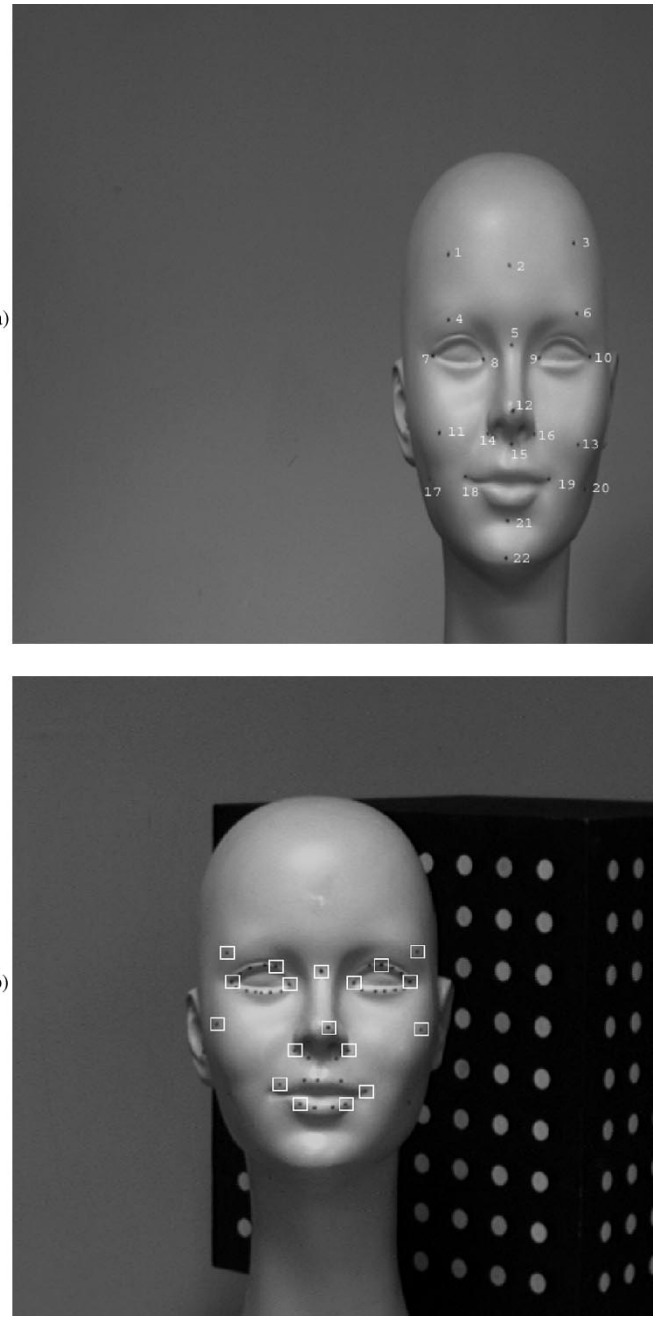


Fig. 1. Reference images of the image sequences. (a) Reference image for tracking by affine transfer, where the labeled dots are the feature points under consideration. (b) Reference image of the target sequence for tracking by orthographic transfer. The feature points overlaid with superimposed squares were selected as reference points.

each of the sequential images. The superimposed images show good overall transfer accuracy. Furthermore, the positions of the points occluded in some images can be predicted in other images. This is demonstrated in Fig. 2(c) where the points at the top of the target head visible in the first image were clipped due to the limited field of view of the camera. However, the missing points are actually predicted in *negative* coordinates in the image and represented by the points in the top row of the image. The four control points for the LCF were assumed to be visible in every image.

TABLE I
TRANSFER ACCURACY OF THE AFFINE TRANSFER METHOD

No.	*Transferred Coordinates*	*Coordinate Measurements*	*Errors*
1	238.43, 62.60	239.00, 62.00	-0.57, +0.60
3	337.46, 53.70	339.00, 54.00	-1.54, -0.30
4	238.62, 114.04	239.00, 114.00	-0.38, +0.04
5	289.21, 134.33	289.00, 134.00	+0.21, +0.33
6	339.64, 110.11	341.00, 110.00	-1.36, +0.11
7	226.83, 142.61	227.00, 143.00	-0.17, -0.39
8	266.46, 144.98	266.00, 145.00	+0.46, -0.02
9	311.04, 144.44	311.00, 144.00	+0.03, +0.44
10	350.66, 143.85	351.00, 143.00	-0.34, +0.85
14	269.65, 204.37	270.00, 204.00	-0.35, +0.37
15	289.49, 212.48	289.00, 213.00	+0.49, -0.52
16	306.30, 204.74	307.00, 205.00	-0.70, -0.26
18	252.93, 237.83	253.00, 239.00	-0.07, -1.17
19	318.32, 240.47	319.00, 240.00	-0.69, +0.47
20	347.08, 249.66	348.00, 249.00	-0.93, +0.66
22	284.86, 302.45	285.00, 303.00	-0.14, -0.55

TABLE II
STATISTICS FOR THE ACCURACY OF THE AFFINE TRANSFER METHOD

Statistics	*Image 2*	*Image 3*	*Image 4*
Arithmetic Mean (pixels)	0.15	-0.17	0.14
Absolute Mean (pixels)	0.53	0.48	0.50
RMSV (pixels)	0.62	0.58	0.61
Absolute Max (pixels)	1.57	1.54	1.45
Absolute Min (pixels)	0.07	0.02	0.03
Subpixel Accuracy (%)	87.5	90.6	87.5

B. Tracking by Orthographic Transfer

Fig. 1(b) shows the reference frame in another image sequence of the target taken by the same camera as Fig. 1(a). The object (cube with an array of white dots on the surfaces) has been placed in the background to create a more complex scene. The 18 feature points with the superimposed squares are manually-selected reference points to be tracked over the image sequence. The second reference frame for the epipolar geometry computation is shown in Fig. 3(b). To verify the transfer performance, we evaluated the edge points detected with the LOG operator in the first reference image. The edge points overlaying the first reference image from Fig. 3(a) were transferred into the other three frames of the target sequence as illustrated in Figs. 3(b), (c) and (d). We see that the overall transfer accuracy in Fig. 3 is comparable to that obtained by the affine transfer method.

To gain a quantitative estimate of the accuracy, we calculated the statistics as in the affine case of the transfer errors of the remaining feature points (dots on the target) over the first three sequential images. Table III lists the statistical results from which

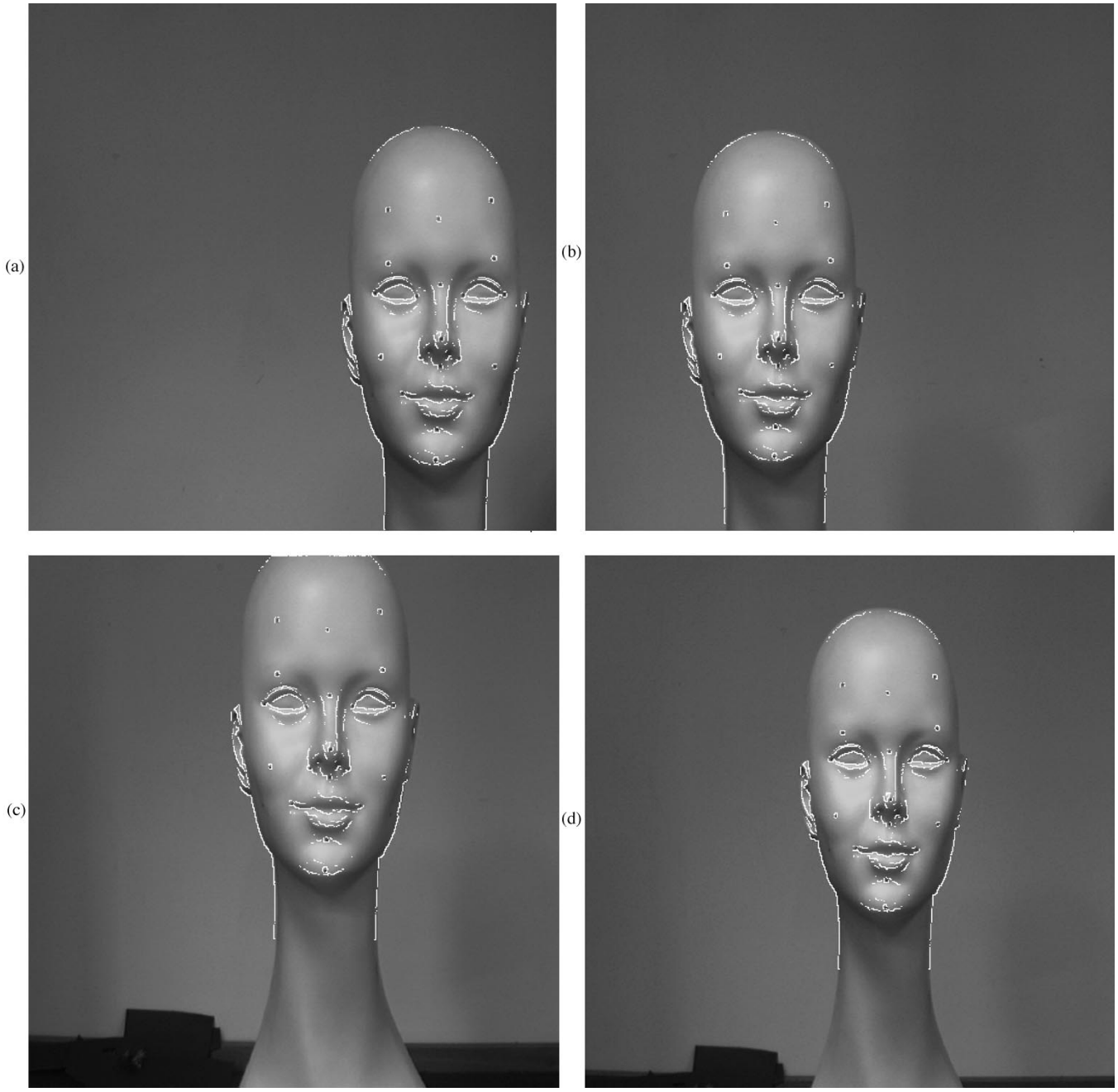


Fig. 2. Feature point transfer over the image sequence, (a), (b), (c), (d). Edge points in the first image were correctly transferred into other images in the sequence.

we can see that orthographic transfer has the same order of accuracy as that obtained by affine transfer. This suggests that subpixel average accuracy can also be achieved using orthographic transfer.

It is not a coincidence that the two methods have the same order of transfer accuracy. Both the methods assume parallel projection models of cameras. The two projection models differ in their projection fashions. There are no constraints on the projection direction from an object to an image plane in the affine case; whereas the projection rays in the orthographic model are always assumed to be perpendicular to an image plane. The similarity of the two projection models can be seen in the two projection functions with similar forms in (5) and (18). Secondly, image transfer is accomplished by a similar epipolar-constrained linear solution by both transfer methods. Therefore, the systemic error, caused by the projection models, and the random errors caused by the measurement and numerical error, should be close for both transfer methods. This result should extend to general sequences.

On the other hand, since the tracking results based on the orthographic transfer method become independent of the positions and configuration of the reference points, a significant improvement in robustness of orthographic transfer has been gained. This is the most important advantage of the orthographic method as compared to the affine transfer method.

Nevertheless, high transfer accuracy was not obtained for the object in the background. The orthographic camera model is valid only when object extents in depth are small compared to

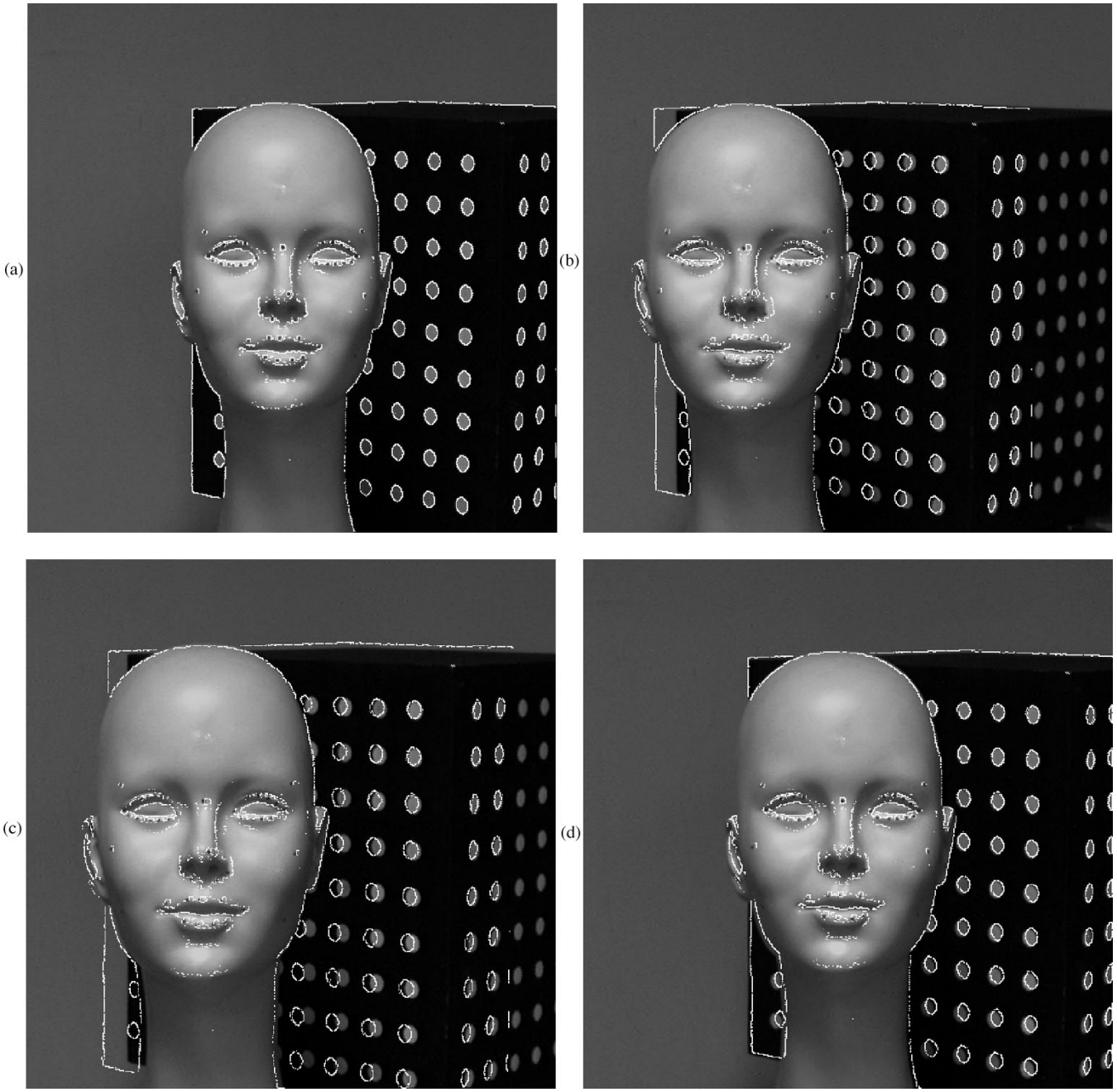


Fig. 3. Edge points transferred from the reference frame into other frames in the target image sequence. The object primarily performs translation over the sequence.

TABLE III
STATISTICS FOR THE ACCURACY OF THE ORTHOGRAPHIC TRANSFER METHOD

Statistics	*Image 2*	*Image 3*	*Image 4*
Arithmetic Mean (pixels)	0.23	0.10	0.01
Absolute Mean (pixels)	0.49	0.51	0.44
RMSV (pixels)	0.62	0.61	0.57
Absolute Max (pixels)	1.37	1.13	1.23
Absolute Min (pixels)	0.00	0.02	0.00
Subpixel Accuracy (%)	84.4	87.5	90.6

the distance between the camera and the object. Therefore, a scene with large depth dimension suffers distortions in orthographic images. Since all reference points used in the image transfer method were selected from the target face, the orthographic camera model achieves its highest projection accuracy there. Other points suffer less projection accuracy, and thus affect the transfer accuracy. Consequently, the transferred edge points, such as those on the left-hand edge of the cube, deviated from their actual positions in the images. This situation can be clearly seen in Figs. 3(b) and (c).

The orthographic transfer method was also applied to an image sequence consisting of 182 frames of a model building.

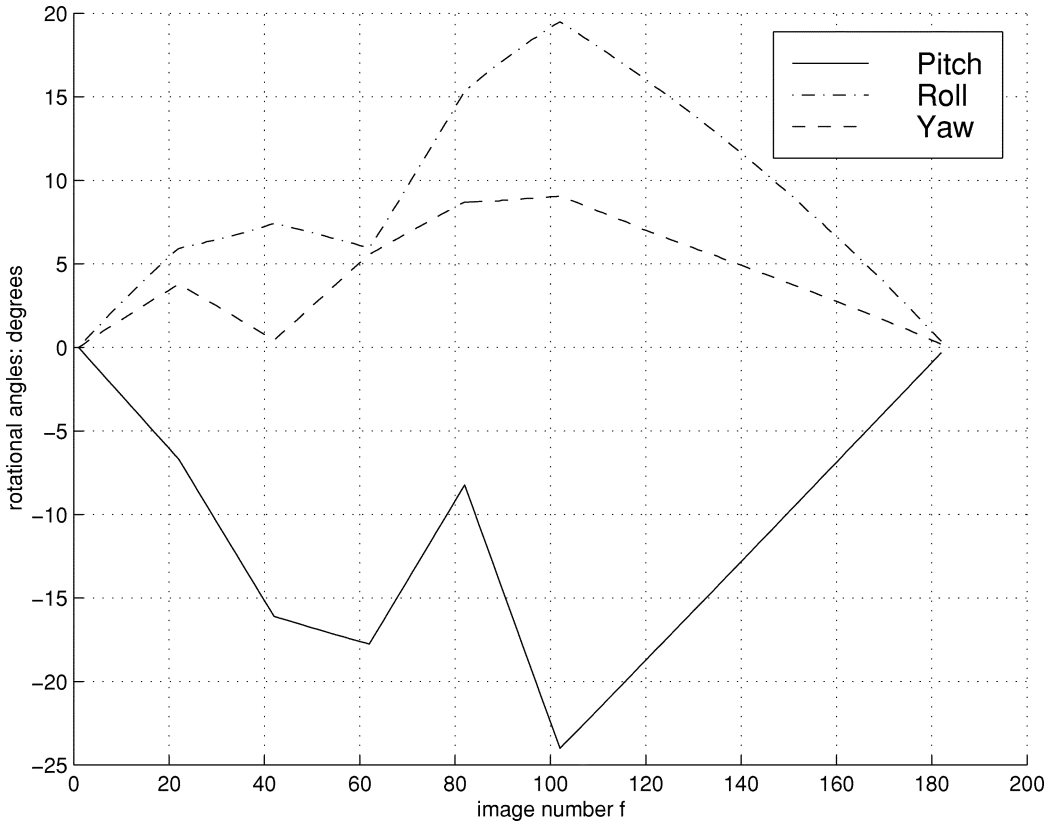


Fig. 4. Actual pitch, roll, and yaw rotations performed by the camera in the image sequence.

The sequence was acquired with a CCD camera performing precise pitch, roll, and yaw motion about the target object. The image size of this sequence was 512×480 pixels. Fig. 4 illustrates the three types of rotational data recorded by the joint encoders of the camera positioning platform.

The first reference frame used for the image transfer experiment was selected as the 100th frame of the image sequence. A set of 32 reference points was selected in the reference frame and manually matched to the corresponding points in other frames. Fig. 5 shows the reference points indicated by small squares in the reference image. The points indicated by cross marks were used to test the transfer accuracy. The second reference frame for the epipolar computation is shown in Fig. 6(b).

Similarly, as in the previous case, the edge points detected in the first reference image of Fig. 5 were used as target points for verifying the performance of the transfer method. The target frames into which the target points are to be transferred were selected as frames 100, 105, 110, 115, 120, and 125 in the image sequence. By referring to Fig. 4 we see that the target frames are within a range for which the three camera rotations are approximately linear. This approach for image selection simplifies the dependence of transfer accuracy on camera rotation. The overall transfer performance can be assessed visually from Fig. 6. The data in Fig. 7 were obtained by calculating the differences between the transferred image coordinates and the manually-selected image coordinates. Statistics for the transfer error shown in Fig. 7 are given in Table IV.

Figs. 6, 7 and Table IV show that the transferred edge points for frames 105 and 110 match their actual positions very closely. The outline of the target was correctly determined and most details of the object were matched satisfactorily in the images. However, the transfer accuracy decreases as the camera rotates from the reference position. This is because the orthographic camera model that is least-calibrated does not simulate the camera well when perspective effects in the images caused by camera rotation are obvious.

Fig. 5. Reference image and reference points of the image sequence.

Since the transfer errors with respect to small camera rotations are small, the performance degradation can be reduced by

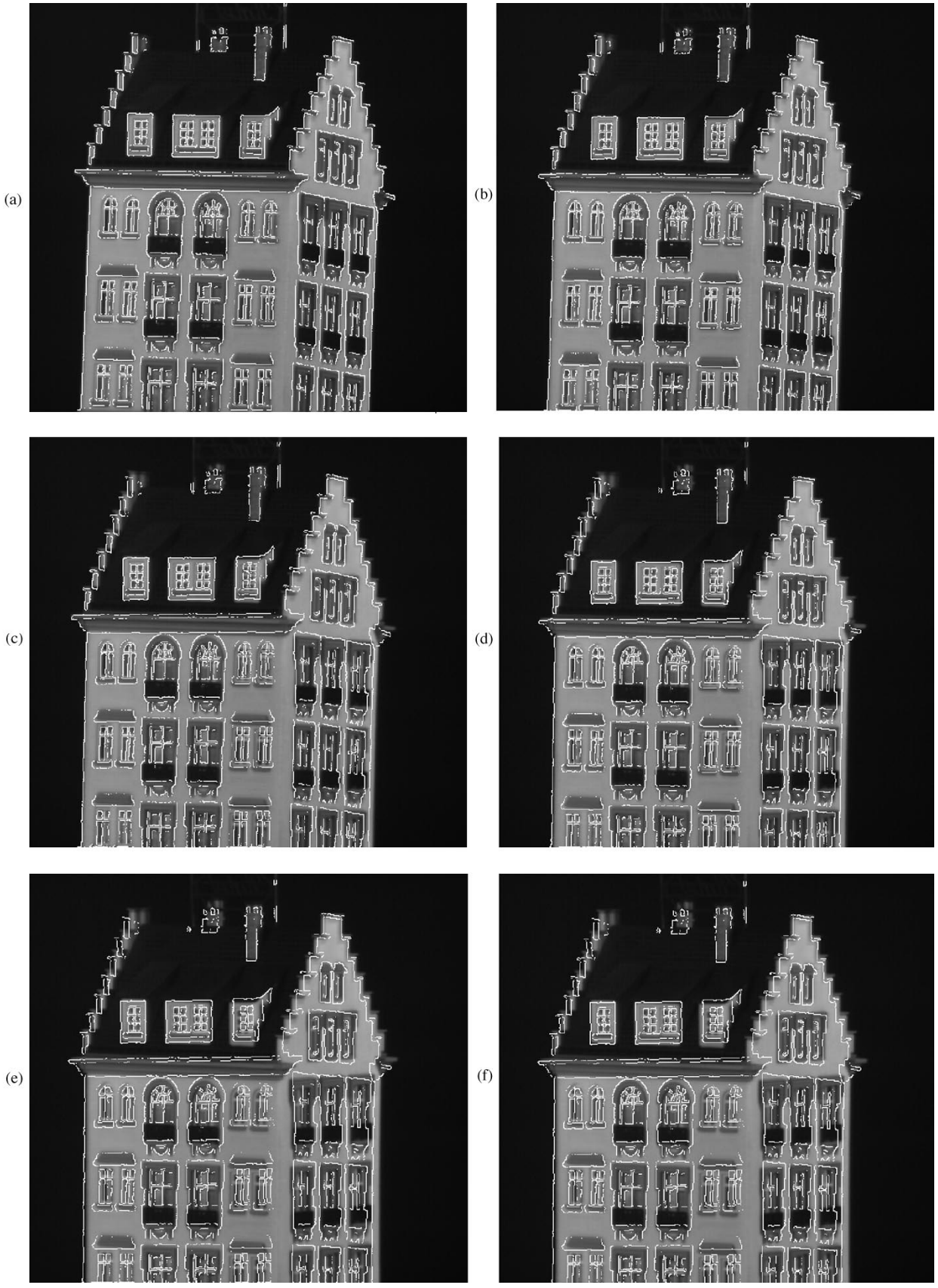


Fig. 6. Edge points transferred from the reference frame into other frames in the image sequence. (a)–(f) correspond to frames 100, 105, 110, 115, 120, and 125 in the sequence, respectively.

a recursive transfer scheme in which the step size of the transfer iteration is limited by a small camera rotation and the output of the current iteration is used as the input for the next iteration. Fig. 8 shows the transfer errors using the recursive scheme,

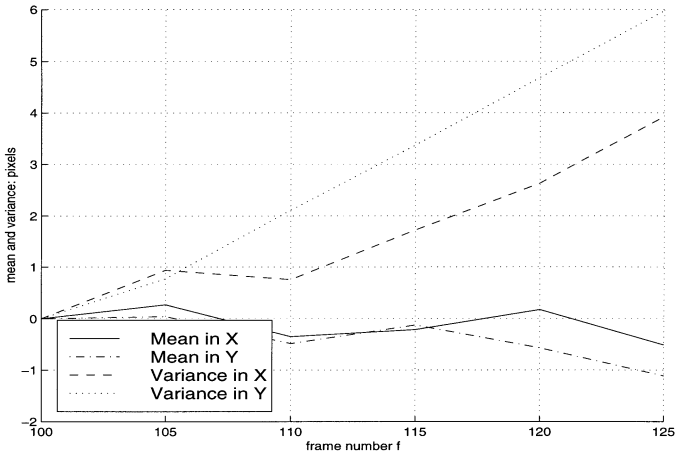


Fig. 7. Transfer errors of the image sequence.

TABLE IV
STATISTICS FOR THE ACCURACY OF THE ORTHOGRAPHIC TRANSFER METHOD ON THE MODEL BUILDING SEQUENCE TRANSFER METHOD

Statistics	*Image 2*	*Image 3*	*Image 4*	*Image 5*	*Image 6*
Arithmetic Mean (pixels)	-0.16	0.42	0.17	0.25	0.81
Absolute Mean (pixels)	0.76	1.27	2.07	3.09	3.90
RMSV (pixels)	0.87	1.59	2.68	3.90	5.07
Absolute Max (pixels)	1.64	4.06	6.36	9.96	13.36
Absolute Min (pixels)	0.18	0.04	0.13	0.12	0.33
Subpixel Accuracy (%)	71.9	43.8	31.3	15.6	25.0

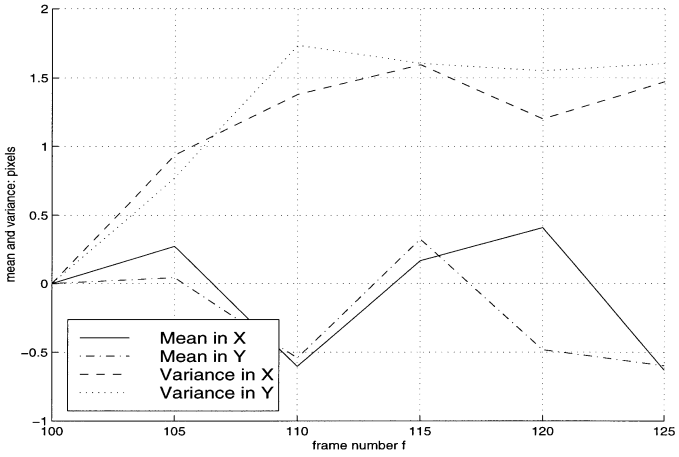


Fig. 8. Transfer errors of the image sequence using the recursive transfer method.

where the iteration step is 5 frames. The transfer accuracy improved significantly when compared to Fig. 7.

V. CONCLUSIONS

We have developed image transfer methods based on affine camera and orthographic camera models. The novel feature of our methods is the elimination of the requirement for target point correspondence over two reference views. The transfer methods were analyzed for cases of novel view synthesis, where trajectories of dense points were required.

For the affine transfer method, image transfer was achieved by calculating the coordinates of the target point in a local coordinate system that is invariant to camera motion. This is an efficient method because only five control points are needed to transfer a target point from one view to another without the need for correspondence of the target point over the views. If any four of the five control points are matched in any other view, the target point can be transferred into the third view with subpixel accuracy. Furthermore, the method can accommodate occlusion during image transfer. Orthographic transfer was introduced to improve the robustness of the affine transfer method to image noise and control point configuration. The orthographic transfer method employs SVD factorization to construct a framework in which the sequential image is represented under orthographic projection. A least-squares method was used to estimate the epipolar geometry between the reference views and derive an epipolar constraint for the image transfer. Our experimental results demonstrated that both methods achieved subpixel transfer accuracy and the orthographic transfer method has high robustness to image noise.

However, relatively poor performance has been demonstrated with both the orthographic and the affine transfer methods when perspective effects are obvious, since the affine and orthographic camera models have insufficient calibration to accommodate perspective effects. Performance could be improved by the use of other camera models that require more calibration such as the weak perspective or the paraperspective [17].

APPENDIX

Lemma 1

$|\boldsymbol{c}| \neq 0$ if there exist at least five non-coplanar points.

Proof: By determinant calculation, we have

$$|\boldsymbol{C}| = -a\,|\boldsymbol{E}_1| - b\,|\boldsymbol{E}_2| \tag{41}$$

where $\boldsymbol{C}$ is the coefficient matrix in (12) and

$$\boldsymbol{E}_1 = \begin{pmatrix} e_{1x} & e_{2x} & e_{3x} \\ e_{1y} & e_{2y} & e_{3y} \\ e'_{1x} & e'_{2x} & e'_{3x} \end{pmatrix}, \quad \boldsymbol{E}_2 = \begin{pmatrix} e_{1x} & e_{2x} & e_{3x} \\ e_{1y} & e_{2y} & e_{3y} \\ e'_{1y} & e'_{2y} & e'_{3y} \end{pmatrix}.$$

Given a set of four general control points, a system of linear equations in the affine epipolar parameters is written from (10) as

$$\left.\begin{aligned} ax'_0 + by'_0 + cx_0 + dy_0 &= -1 \\ ax'_1 + by'_1 + cx_1 + dy_1 &= -1 \\ ax'_2 + by'_2 + cx_2 + dy_2 &= -1 \\ ax'_3 + by'_3 + cx_3 + dy_3 &= -1 \end{aligned}\right\}. \tag{42}$$

Assume that this set of control points is identical to that for the affine coordinate computation. Then, subtracting the first line in (42) from the second and the third line from the fourth, yields the following simultaneous equations:

$$\left.\begin{aligned} ae'_{1x} + be'_{1y} + ce_{1x} + de_{1y} &= 0 \\ ae'_{2x} + be'_{2y} + ce_{2x} + de_{2y} &= 0 \\ ae'_{3x} + be'_{3y} + ce_{3x} + de_{3y} &= 0 \end{aligned}\right\}. \tag{43}$$

Rewriting (43) in matrix form, we have

$$\begin{pmatrix} e'_{1x} & e'_{1y} & e_{1x} \\ e'_{2x} & e'_{2y} & e_{2x} \\ e'_{3x} & e'_{3y} & e_{3x} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = -d \begin{pmatrix} e_{1y} \\ e_{2y} \\ e_{3y} \end{pmatrix}. \tag{44}$$

Since the basis $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3\}$ has linearly independent vectors, the images of $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3\}$ in the two affine views are also linearly independent, which yields:

$$|\boldsymbol{E}| = \begin{vmatrix} e'_{1x} & e'_{1y} & e_{1x} \\ e'_{2x} & e'_{2y} & e_{2x} \\ e'_{3x} & e'_{3y} & e_{3x} \end{vmatrix} \neq 0.$$

Then, a and b are computed by

$$a = \frac{|\boldsymbol{E}_a|}{|\boldsymbol{E}|}, \qquad b = \frac{|\boldsymbol{E}_b|}{|\boldsymbol{E}|} \tag{45}$$

where

$$|\boldsymbol{E}_a| = -d \begin{vmatrix} e_{1y} & e'_{1y} & e_{1x} \\ e_{2y} & e'_{2y} & e_{2x} \\ e_{3y} & e'_{3y} & e_{3x} \end{vmatrix}$$

$$|\boldsymbol{E}_b| = -d \begin{vmatrix} e'_{1x} & e_{1y} & e_{1x} \\ e'_{2x} & e_{2y} & e_{2x} \\ e'_{3x} & e_{3y} & e_{3x} \end{vmatrix}.$$

By substituting (45) into (41) we have

$$\begin{aligned} |\boldsymbol{C}| &= -\frac{d}{|\boldsymbol{E}|}\left(-|\boldsymbol{E}_1|\,|\boldsymbol{E}_a| - |\boldsymbol{E}_2|\,|\boldsymbol{E}_b|\right) \\ &= -\frac{d}{|\boldsymbol{E}|}\left(\left|\boldsymbol{E}'_1\right|\,|\boldsymbol{E}_a| - \left|\boldsymbol{E}'_2\right|\,|\boldsymbol{E}_b|\right) \\ &= -\frac{d}{|\boldsymbol{E}|}\left(|\boldsymbol{E}_b|\,|\boldsymbol{E}_a| - |\boldsymbol{E}_a|\,|\boldsymbol{E}_b|\right) \\ &= 0. \end{aligned} \tag{46}$$

This suggests that $|\boldsymbol{C}|$ will be zero if the same set of four control points is applied to both the affine basis construction and the affine epipolar parameter computation. To guarantee $|\boldsymbol{C}| \neq 0$, there must be at least one differing element in the two sets of control points for basis construction and epipolar parameter computation. Then, at least *five* points, with no more than four coplanar, are necessary to ensure $|\boldsymbol{C}| \neq 0$.

REFERENCES

[1] S. Avidan and A. Shashua, "Novel view synthesis by cascading trilinear tensors," *IEEE Trans. Pattern Analy. Machine Intelligence*, vol. 4, no. 4, pp. 293–306, Apr. 1998.

[2] O. Faugeras and L. Robert, "What can two images tell us about a third one?," *Int'l. J. Comput. Vision*, vol. 18, no. 1, pp. 5–19, Apr. 1996.

[3] G. H. Golub and C. F. van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, 1996.

[4] E. Hayman, I. Reid, and D. Murray, "Zooming while tracking using affine transfer," in Proc. British Machine Vision Conf., Edinburgh, Scotland, U.K., 1996.

[5] R. Hartley, "Lines and points in three views and the trifocal tensor," *Int'l. J. Comput. Vision*, vol. 22, no. 2, pp. 125–140, Mar. 1997.

[6] A. Heyden, "A common framework for multiple view tensors," in *Proc. Fifth European Conf. Comput. Vision*, Freiburg, Germany, June 1998, pp. 3–19.

[7] F. Kahl and A. Heyden, "Affine structure and motion from points, lines and conics," *Int'l. J. Comput. Vision*, vol. 33, no. 3, pp. 163–180, Sept. 1999.

[8] J. J. Koenderink and A. J. van Doorn, "Affine structure from motion," *J. Optical Society of America*, vol. 8, no. 2, pp. 377–385, Feb. 1991.

[9] P. R. S. Mendonca and R. Cipolla, "Analysis and computation of an affine trifocal tensor," in Proc. 9th Bristish Machine Vision Conf., Southampton, U.K., 1998.

[10] J. L. Mundy and A. Zisserman, *Geometric Invariance in Computer Vision*, J. L. Mundy and A. Zisserman, Eds. Cambridge, MA: MIT Press, 1992.

[11] S. Pollard *et al.*, "View synthesis by trinocular edge matching and transfer," in Proc. 9th Bristish Machine Vision Conf., Southampton, U.K., 1998.

[12] W. H. Press *et al.*, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge University Press, 1996.

[13] L. Quan, "Invariants of six points and projective reconstruction from three uncalibrated images," *IEEE Trans. Pattern Analy. Machine Intelligence*, vol. 17, pp. 34–46, Jan. 1995.

[14] ——, "Self-calibration of an affine camera from multiple views," *Int'l. J. Comput. Vision*, vol. 19, no. 1, pp. 93–105, 1996.

[15] L. Quan and T. Kanade, "Affine structure from line correspondences with uncalibrated affine cameras," *IEEE Trans. Pattern Analy. Machine Intelligence*, vol. 19, pp. 834–845, Aug. 1997.

[16] I. D. Reid and D. W. Murray, "Active tracking of foveated feature clusters using affine structure," *Int'l. J. Computer Vision*, vol. 18, no. 1, pp. 41–60, 1996.

[17] L. S. Shapiro, *Affine Analysis of Image Sequence*. Cambridge, U.K.: Cambridge University Press, 1995.

[18] A. Shashua, "Algebraic function for recognition," *IEEE Trans. Pattern Analy. Machine Intelligence*, vol. 17, pp. 779–789, Aug. 1995.

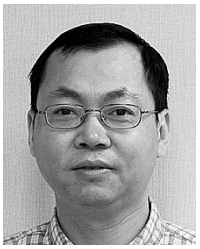
[19] ——, "Trilinear tensor: The fundamental construct of multiple-view geometry and its applications," in Proc. Int'l. Workshop Algebraic Frames for the Perception Action Cycle, Kiel, Germany, 1997.

[20] A. Shashua and L. Wolf, "Homography tensors: On algebraic entities that represent three views of static or moving planar points," in Proc. 6th European Conf. Computer Vision, Dublin, Ireland, 2000.

[21] T. Thorhallsson and D. Murray, "The tensors of three affine views," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition*, Fort Collins, CO, June 1999, pp. 450–456.

[22] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int'l. J. Comput. Vision*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[23] S. Ullman and R. Basri, "Recognition by linear combination of models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 992–1006, Oct. 1991.



Jason Z. Zhang (M'97) received the B.Sc. degree in electronic engineering from Nanjing University of Posts and Telecom, the M.Sc. degree in communications and electronic systems from Beijing University of Posts and Telecom, and the Ph.D. degree in signal and information processing from Northern Jiaotong University, China, in 1984, 1989, 1994 respectively.

He is a Research Officer in Vancouver Innovation Centre of National Research Council of Canada, where his current research interests include geometric vision methods for video processing and understanding, 3D microscopic imaging, and characterization of the microstructural morphology of the materials and components in fuel cell systems. Prior to joining NRC in 1999, he was a Research Assistant with the Department of Electronic Engineering of The Chinese University of Hong Kong (1996–1997) and a Post Doctoral Fellow with Centre for Signal Processing of Nanyang Technological University of Singapore (1997–1999). His research background consists of multi-view environment modeling and understanding, active vision for surveillance and robot guidance and navigation, video object extraction and representation, real-time multi-video-stream compression and communication and 3D imaging and analysis of microstructural morphology of materials.

Q. M. Jonathan Wu (M'92) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990.

He is a Research Officer and Group Leader at the Vancouver Innovation Center, National Research Council Canada, Vancouver, BC, Canada. He is also an Adjunct Professor in the School of Engineering Science, Simon Fraser University. From 1982 to 1984, he was a Lecturer at Shandong University. From 1992 to 1994, he was with the University of British Columbia as a Senior Research Scientist. His research interests include pattern recognition, image analysis, neural networks, robotics, intelligent control and computer vision systems.

Dr. Wu was a recipient of the Visiting Fellowship of the BC Advanced Systems Institute. He received one of the major awards in the Control and Automation field—the Best Control Applications Paper Prize, awarded at the 11th IFAC World Congress held in Tallinn in 1990.

Hung-Tat Tsui (M'82) received the B.Sc.(Eng) degree in electrical engineering from the University of Hong Kong in 1964, the M.Sc. degree from the University of Manchester Institute of Science and Technology in 1965, and the Ph.D. degree from the University of Birmingham in 1969.

He joined the Mathematical section of the Central Electricity Research Laboratories at Leatherhead, U.K. as a research officer in April 1969. In December 1971, he joined the Department of Electronics (later becomes the Department of Electronic Engineering) of the Chinese University of Hong Kong where he is now a professor. His current interests include computer vision, 3D reconstruction using a hand held camera, image base rendering, vision systems for mobile robots, 3D ultrasound imaging, image guided surgery, modeling, and animation of human faces. He has published more than 80 refereed papers in international journals and conference proceedings in the above areas in the past 15 years.

William A. Gruver (M'95) received the B.S.E.E, M.S.E.E and Ph.D. degrees in electrical engineering from the University of Pennsylvania in 1963, 1966 and 1970, respectively; and the DIC in Automatic Control Systems from Imperial College of Science and Technology, London, in 1965.

He is Professor of Engineering Science at Simon Fraser University in Burnaby, British Columbia, Canada. His industrial experience includes management and technical leadership positions at GE's Factory Automation Products Division in Charlottesville; GE Automation Center in Frankfurt, Germany; IRT Corporation in San Diego, the Center for Robotics and Manufacturing Systems at the University of Kentucky and LTI Robotic Systems, a California-based startup that he co-founded. He has also held engineering and faculty positions at NASA's Marshall Space Flight Center, DFVLR German Space Research Center, Technical University Darmstadt, U.S. Naval Academy and North Carolina State University.

He has published 175 technical articles and three books on robotics, automation, control and optimization. His current research focus is intelligent distributed systems for industrial automation and resource management with applications to manufacturing and energy systems. In 1995, Dr. Gruver was elected an IEEE Fellow for his "leadership of major international programs in robotics and manufacturing systems engineering." Dr. Gruver is an Associate Editor of the IEEE Transactions on SMC and he chairs the technical committee on Robotics and Manufacturing Automation of the IEEE SMC Society. Currently, he is Vice President of Finance and Long Range Planning of the IEEE SMC Society and was previously VP Publications and VP Conferences. He has served as associate editor for the IEEE Transactions on Robotics and Automation, associate editor of the IEEE Transactions on Control Systems Technology and a founding officer of the IEEE Robotics and Automation Society. He has served as General Chair of the 1994 IEEE International Conference on Robotics and Automation, General Chair of the 1995 IEEE International Conference on SMC, Program Co-Chair of the 1999 IEEE International Conference on SMC and General Co-Chair of the 2001 Joint 9th IFSA World Congress and 20th NAFIPS International Conference.