Time-varying Shape Tensors for Scenes with Multiply Moving Points

Anat Levin Lior Wolf Amnon Shashua*

School of Computer Science and Engineering,
The Hebrew University,
Jerusalem 91904, Israel
http://www.cs.huji.ac.il/$\sim$ $\{\mathrm{alevin}, \mathrm{lwolf}, \mathrm{shashua}\}$/

Abstract

We derive single view indexing functions for dynamic scenes — where dynamic is defined as a scene consisting of multiply moving points each moving independently with constant velocity. The indexing functions we derive are view independent and form a generalization of the "shape tensors" associated with rigid scenes by introducing a time-varying parameter. We derive those indexing functions under full 3D projective, 3D affine, and various reduced configurations.

The indexing functions were implemented and tested for matching against objects for which their non-rigid motion is an intrinsic part of their character — human gait recognition and hand gesture identification are the two chosen application examples.

1 Introduction

In the context of multi-view analysis of 3D rigid scenes two kinds of algebraic invariants are often discussed and elaborated upon — one is invariant to the 3D shape of the object (multi-view tensors such as the fundamental matrix of two views and the trilinear and quadlinear tensors of 3,4 views respectively [18, 10, 21, 13, 19]) and the other is invariant to the viewing geometry (shape tensors). The latter is often used in the context of recognition of rigid objects under varying viewing positions and is based on the following principle. A 3D rigid object is modeled as a point configuration projected onto multiple 2D views. The projection is a function of the 3D coordinates, the camera position, and the image coordinates. By having sufficiently many points in a single view, the camera parameters can be algebraically eliminated leaving constraints which involve the 3D coordinates and the 2D coordinates alone. These constraints are known as *single view shape constraints* and have appeared in [23, 6, 17, 9, 22, 14, 15, 8].

In this paper we derive single view shape constraints of a continuously changing 3D shape — i.e., the notion of *time* is integrated into the shape constraints. The continuous change model is inspired by the recent work on motion understanding of 3D *dynamic* point configurations, where dynamic is defined in the sense of having multiply moving points along straight line (or curved) paths (and in some case with constant velocity) [2, 20, 16, 24, 12, 25]. The non-rigidity in this framework is "structured" in a way which is amenable to algebraic treatment — for example, multi-view tensors for planar scenes [20] and general 3D scenes [12, 25] were introduced for dynamic point configurations and which form a natural generalization of the classical multi-view tensors associated with rigid configurations.

In the same vain, we wish to generalize the shape tensors introduced in [23, 6] which were associated with rigid point configurations to the case of dynamic configurations. Unlike the rigid case where the shape tensors are a function of image measurements and 3D point positions, here the notion of time is represented as well (since the point configuration changes relative positions as a function of time). We will derive the single view "shape+motion" invariants starting from the 3D affine projection model, then the full 3D projective (perspective) model, then consider reduced configurations and apply these models to the case of multiply moving points along straight line paths with constant velocity and second-order paths (the latter only for the full 3D projective model).

We have implemented and tested the indexing functions and applied them to the task of matching against classes of objects characterized both by their shape and motion. The classes we selected in this paper include human gait recognition, people making a "sitting" motion (from upright position to a full sitting position), and hand gesture recognition. Using the indexing function we were able to match a single view of (a person walking, sitting, or hand gesture) to the correct class, i.e., is this an image of a person walking? sitting? or what class of hand gesture is it? Although the experiments are not intended to introduce a complete system for hand gesture recognition or human gait classification, they do demonstrate the relevance of the dynamic indexing functions we derive to real applications of interest.

*A.S. is on sabbatical at the CS department of Stanford university and can be reached at shashua@cs.stanford.edu

1.1 Related Work

The closest work to ours is the use of view-consistency constraints for human gait recognition by matching image sequences [7]. The classical view consistency constraint, also known as "recognition polynomials" [3, 4] is based on the following principle. Recall that the shape tensors arise from an elimination process in which given sufficiently many points in a single view, the camera parameters can be algebraically eliminated leaving constraints which involve the 3D coordinates and the 2D coordinates alone. The process of elimination can continue by having a number of views until we are left with constraints involving image coordinates alone. These constraints, known as *view consistency constraints*, express the fact that those number of views of those number of points are the projections of the same 3D object. The second step of elimination is unwieldy under general projective setting and has therefore not received much attention, however, is relatively manageable under orthographic or scaled orthographic projection. Under orthographic projection two views of four points are sufficient for a constraint which was first derived in [3, 4]. For scaled orthographic projection, two views of five points are necessary for a constraint which was recently derived in [7].

The 5-point view consistency constraint of [7] was introduced for human gait recognition by comparing two sequences of the same person walking — each sequence from a different viewing position. The view consistency constraint was evaluated for every pair of frames, one from each sequence, thus creating a "sequence consistency matrix" whose entries are the residual of the view consistency constraint (low residual reflects good consistency). Thus, the diagonal of the matrix should have low numbers if indeed the two sequences are of the same person. In other words, the role of a sequence in this approach is to add a statistical component on top of the basic two-view algebraic consistency expression. Thus, this approach is fundamentally different from the task set out in this paper which is to define a new single-view shape contraint of "structured" non-rigid phenomena, in which the structuring takes the form of multiply moving points along straight-line (or second order curves) paths with constant velocity.

2 Space-time Single-view Constraints

Consider a 3D point configuration $P_1, ...P_n$ which move along straight line paths with constant velocity $V_1, ..V_n$. The position of the i'th point at time $j = 0, 1, ..., m$ is $P_i + jV_i$. Let M_j denote the 3×4 camera matrix projecting all points $P_i + jV_i$ onto the j'th image plane, and let $p_{ij} = (x_{ij}, y_{ij}, 1)^\top$ be the projection of $P_i + jV_i$ at view j, thus we have:

$$p_{ij} \cong [M_j] \begin{pmatrix} P_i + jV_i \\ 1 \end{pmatrix} \quad (1)$$

Our goal is that given a sufficient number of image points in a single view to algebraically eliminate the camera parameters leaving constraints which involve the 3D coordinates P_i, the velocity vectors V_i, powers of the time parameter j and the image coordinates. We will start doing so for the 3D affine camera model and then proceed to the general perspective model.

2.1 Affine Cameras

For the affine camera model (projection rays are parallel and meet the image plane at some oblique angle), the camera matrix has the following form:

$$M_j = \begin{bmatrix} \boldsymbol{a}_j^\top & r_j \\ \boldsymbol{b}_j & s_j \\ 0^\top & 1 \end{bmatrix}_{3\times 4}$$

The projection equation 1 can be further manipulated:

$$\begin{bmatrix} P_i^\top + jV_i^\top & 1 & x_{ij} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_j \\ r_j \\ 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} P_i^\top + jV_i^\top & 1 & y_{ij} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_j \\ s_j \\ 1 \end{bmatrix} = 0$$

Given 5 such points, we take A, B to be:

$$A = \begin{bmatrix} P_1^\top + jV_1^\top & 1 & x_{1j} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ P_5^\top + jV_5^\top & 1 & x_{5j} \end{bmatrix}_{5\times 5}$$

$$B = \begin{bmatrix} P_1^\top + jV_1^\top & 1 & y_{1j} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ P_5^\top + jV_5^\top & 1 & y_{5j} \end{bmatrix}_{5\times 5}$$

Since:

$$A \begin{bmatrix} \boldsymbol{a}_j \\ r_j \\ 1 \end{bmatrix} = 0$$

$$B \begin{bmatrix} \boldsymbol{b}_j \\ s_j \\ 1 \end{bmatrix} = 0$$

We conclude that $det(A) = 0$ and $det(B) = 0$. The determinant expansions of $det(A)$ and $det(B)$ are polynomials $f_1(p_{1j}, ..., p_{5j}, j)$ and $f_2(p_{1j}, ..., p_{5j}, j)$ in the five image points and j (the time parameter), whose coefficients are functions of the (maybe unknown) shape and direction parameters P_i and V_i.

We can simplify the above determinants by selecting a 2D affine canonical basis for the $j'th$ images plane where the first 3 points are $p_{1j} = (0,0,1)^\top$, $p_{2j} = (1,0,1)^\top$, and $p_{3j} = (0,1,1)^\top$. Thus, the polynomials $f_1(), f_2()$ are functions of the 4'th and 5'th image points and the parameter j alone. For example, the matrix A has the form:

$$A = \begin{bmatrix} P_1^\top + jV_1^\top & 1 & 0 \\ P_2^\top + jV_2^\top & 1 & 1 \\ P_3^\top + jV_3^\top & 1 & 0 \\ P_4^\top + jV_4^\top & 1 & x_{4j} \\ P_5^\top + jV_5^\top & 1 & x_{5j} \end{bmatrix}_{5\times 5} \quad (2)$$

and $f_1(x_{4j}, x_{5j}, j) = det(A)$ is linear in the image coordinates x_{4j}, x_{5j} and order 3 in j. Therefore the non-vanishing terms contain the elements of the Cartesian product $[x_{4j}, x_{5j}, 1] \otimes [1, j, j^2, j^3]$ and, likewise, the non vanishing terms in $\det(B)$ are elements of $[y_{4j}, y_{5j}, 1] \otimes [1, j, j^2, j^3]$.

This implies that $f_1(), f_2()$ have 12 non vanishing coefficients each. However, since the coefficients of x_{5j} is identical to that of y_{5j} (which is $det(A_{1-4,1-4})$) — and likewise the coefficient of x_{4j} is identical to that of y_{4j} — there are only 16 distinct coefficients for f_1, f_2 together. Since each image frame provides 2 constraints ($f_1(x_{4j}, x_{5j}, j) = 0$ and $f_2(y_{4j}, y_{5j}, j) = 0$), then 8 frames suffice to solve for the unknown coefficients of the polynomials f_1, f_2. We summarize this in the claim below:

Claim 1 *A single view of 5 general dynamic points $P_i + jV_i$ provide two affine invariants which are linear in the image coordinates and of order 3 in the time parameter j. The two invariants together have 16 distinct view-independent coefficients which are a function of P_i, V_i and which can be recovered linearly from 8 views.*

The process of single-view indexing proceeds as follows. Given 8 views of the configuration of 5 dynamic points, the 16 view-independent coefficients can be recovered linearly. For any novel view of the configuration $P_i + jV_i$, for some unknown j, the functions $f_1(), f_2()$ are polynomials of order 3 in j, i.e., $\alpha_0 + \alpha_1 j + \alpha_2 j^2 + \alpha_3 j^3 = 0$ (with known coefficients $\alpha_0, ..., \alpha_3$). Since we have two polynomials in j from each group of 5 points taken from the object, a minimum of two distinct groups of 5 points would be sufficient for (linear) consistency verification (intersection of planes in 3D). A necessary condition for the image of these point-configurations to belong to the object class $P_i + jV_i$ is that the collections of planes (two for each group of 5 points) intersect at a single point, i.e., there is a consistent time parameter j at the intersection. A single configuration (two polynomials) may also be sufficient by intersecting the finite solutions for j from each polynomial — a necessary condition for the configuration to belong to the class is that the intersection is not null.

2.2 General Perspective Camera

Generalizing the affine space-time view-independent functions introduced above to full projective requires a minimum of 6 points, as follows. Let l_{ij}, l'_{ij} in the j'th image be lines coincident with the point p_{ij}, and denote by A the matrix:

$$A = \begin{bmatrix} l_{1j} \otimes [P_1 + jV_1, 1]^\top \\ l'_{1j} \otimes [P_1 + jV_1, 1]^\top \\ \cdot \\ \cdot \\ l_{6j} \otimes [P_6 + jV_6, 1]^\top \\ l'_{6j} \otimes [P_6 + jV_6, 1]^\top \end{bmatrix}_{12\times 12}$$

The determinant of A vanishes since:

$$A_{12\times 12} \begin{bmatrix} m_{1j}^\top \\ m_{2j}^\top \\ m_{3j}^\top \end{bmatrix}_{12\times 1} = 0,$$

where m_{1j}, m_{2j}, m_{3j} are the rows of the camera matrix M_j. We can simplify the determinant expansion by selecting a canonical basis for the first four image points $p_{1j} = (0,0,1)^\top$, $p_{2j} = (0,1,0)^\top$, $p_{3j} = (1,0,0)^\top$ and $p_{4j} = (1,1,1)^\top$ leaving a function $f(p_{5j}, p_{6j}, \hat{j}) = 0$ where $\hat{j} = [1, j, ..., j^9]$ (note that j appears only in 9 of the columns of A) which is bilinear in the 5'th and 6'th image points and of order 9 in the time parameter j. Thus, the constraint, viewed as a tensor, has $3 \times 3 \times 10$ elements. The contraction of this tensor $\mathcal{T}^{abc}$ with p_{5j}, p_{6j} and $\hat{j} = [1, j, ..., j^9]$ vanishes whenever those points and time index arise from the desired configuration.

The 90 elements of the tensor are not independent as there are 40 linear constraints among the elements of the tensor. These constraints arise as follows. Let $e_1 = (1,0,0), ..., e_4 = (1,1,1)$ denote the standard basis. Consider the camera matrix whose i'th column consists of e_i and all the remaining entries vanish. Then, M maps the 3D projective space either to $(0,0,0)$ or to e_i, therefore, regardless of the positions of the 3D point configuration and their velocities, $p_{5j} \cong p_{6j} \cong e_i$ (or they vanish), thus $det(A) = 0$. Therefore, for any 3D configuration and for any time j, the contraction of the tensor with $e_i, e_i, [1, j, ..., j^9]$ (where e_i is one of the 4 basis points) must vanish. Since this is true for all j we can conclude that the contraction of the tensor with e_i, e_i would give us a vector of ten zeros. As a result, the tensor contains only 50 parameters (up to scale). Since each view provides one (linear) constraint for the 50 parameters, we will need 49 views (of the 6 dynamic points) in order to recover the shape tensor. This is summarized below:

Claim 2 *In a projective frame, a single view of 6 dynamic points $P_i + jV_i$ provide a projective view-independent invariant which is bilinear in the image coordinates and of order 9 in the time parameter j. The invariant function is a $3 \times 3 \times 10$ tensor whose contraction with the 5'th and 6'th image points and the vector $(1, j, ..., j^9)$ vanishes. The tensor is defined by 50 parameters up to scale (90 entries of the tensor minus 40 linear constraints among the tensor elements), thus 49 views are required for computing the invariant function.*

The process of single-view indexing proceeds in the same manner as in the affine case. For every configuration of 6 points, one can recover the view-independent tensor (from 49 views). For any novel view of each such configuration, the contraction with the tensor yields a 9'th degree polynomial in the time parameter j. Thus with at least two such configurations one can intersect the solutions for j — a necessary condition that the configurations of points arise from $P + i + jV_i$ is that the intersection is not null.

The constraints above, however, are not stable numerically as they contain large exponents of the frame number j. In many practical situations the non-rigidity can be described by motion which is less general then the 3D constant velocity model. In these cases we are able to derive smaller tensors which are more numerically stable to compute. Simplifications can be made by limiting the generality of the position of 3D points at one hand, or by limiting the generality of the directions of the motion on the other hand. Some examples are described below.

2.2.1 Coplanar Trajectories

In case $\dim span\{V_i\} = 2$, we can assume up to affine transformation that all the velocities lie in the XY plane. i.e The Z component of V_i vanishes. Let A be as defined in equation 2.2. The time index j appears only in 6 of the columns of A. The highest exponent of j in the determinant expansion of A becomes 6. The size of the resulting tensor in this case would be $3^2 \times 7 = 63$ out of which 35 are linearly independent (i.e., 34 views are necessary to compute the model).

2.2.2 Coplanar points, 3D velocities

Consider the case where at least on one time along the action, all the points become coplanar. In this case we can assume (up to affine transformation) that the third coordinate of each point P_i to vanish. Three columns of A now contain only multiplications of j. The exponent of j would rise from 3 to 9, as j would appear at least three times in every monomial of the determinant expansion. Since j^3 is shared among all the factors, we can view j^3 as a common scale factor of our tensor. So the resulting constraint would contain only the powers of j from zero to six. The size of the resulting tensor would be 63. Of these 63 elements only 35 would be independent.

2.2.3 Coplanar points, Collinear velocities

Assume that, in addition to the above, all points move in the same direction, which is not contained in the plane of the point configuration. Up to affine transformation, we can assume that this direction is along the Z axis. Thus, only 3 of the columns of A contain multiplications with j (the 3 columns matching the Z coordinate), and they contains only multiplications with j. Since j^3 is shared among all the factors, we can view j^3 as a common scale factor of our tensor, which does not affect the vanishing of the determinant. Therefore the tensor derived in this case is *time independent*, and contains only 3×3 elements (Of these 9 elements only 5 would be independent).

To summarize, the tensor size corresponding to the various reduced configurations are displayed in the table below.

	$\dim span(P_i) = 3$	$\dim span(P_i) = 2$
$\dim span(V_i) = 3$	$3^2 \times 10$	$3^2 \times 7$
$\dim span(V_i) = 2$	$3^2 \times 7$	$3^2 \times 4$
$\dim span(V_i) = 1$	$3^2 \times 4$	3^2

3 Motion Along Constrained Elliptic Trajectories

We have discussed so far view-invariant polynomials for dynamic scenes which contain multiply moving points along straight-line paths with constant velocity. It is possible to obtain simple view-invariant polynomials also for some constrained non-linear motions, such as multiply moving points along elliptic trajectories with constant angular velocity.

Let P_i be 3D points moving along elliptic trajectories. Let the centers of the trajectories be O_i, and let the axes of the ellipse be U_i, V_i. At time j, the 3D location of the point is $P_i = O_i + \cos(\lambda j)U_i + \sin(\lambda j)V_i$.

Assume that all the points are coplanar at some time along the motion, i.e., w.l.o.g the third coordinate of P_i vanishes for all i. Also assume that one of the axes of each ellipse is perpendicular to the XY plane, i.e., $U_i = (0, 0, u_i)$ and $V_i = (v_i, 0, 0)$. In this case the matrix A of equation 2.2

becomes:

$$A = \begin{bmatrix} l_1 \otimes [X_1 + \sin(\lambda j)v_1, Y_1, \cos(\lambda j)u_1, 1]^\top \\ l'_{1j} \otimes [X_1 + \sin(\lambda j)v_1, Y_1, \cos(\lambda j)u_1, 1]^\top \\ . \\ . \\ l_{6j} \otimes [X_6 + \sin(\lambda j)v_6, Y_6, \cos(\lambda j)u_6, 1]^\top \\ l'_{6j} \otimes [X_6 + \sin(\lambda j)v_6, Y_6, \cos(\lambda j)u_6, 1]^\top \end{bmatrix}_{12\times 12}$$

The determinant of A vanishes since:

$$A_{12\times 12} \begin{bmatrix} m_{1j} \\ m_{2j} \\ m_{3j} \end{bmatrix}_{12\times 1} = 0.$$

Notice that A has six columns which do not contain any function of j. Three columns of A contain only multiplications of $\cos(\lambda j)$, and therefore $\cos^3(\lambda j)$ can be seen as a general scale factor. Another three columns contain multiplications of $\sin(\lambda j)$ along with other elements (the first coordinate of O_i) . So the constraint which we get from the determinant expansion of the matrix A, is multilinear in the fifth, and six image points and a power series of $\sin(\lambda j) : 1, \sin(\lambda j), \sin^2(\lambda j), \sin^3(\lambda j)$.

One can reduce the type of elliptic motion even further and thus obtain a simpler invariant — this time both view-independent and time-independent. Assume that in addition to the above there is a known constant ratio δ between the axis of the motion ellipse. i.e for every i, there exist α_i such that $V_i = (\alpha_i \quad 0 \quad 0)^\top$ and $U_i = (0 \quad 0 \quad \delta\alpha_i)^\top$. In this case the matrix A takes the form:

$$A = \begin{bmatrix} l_1 \otimes [X_1 + \sin(\lambda t)\alpha_1, Y_1, \cos(\lambda t)\delta\alpha_1, 1]^\top \\ l'_{1j} \otimes [X_1 + \sin(\lambda t)\alpha_1, Y_1, \cos(\lambda t)\delta\alpha_1, 1]^\top \\ . \\ . \\ l_{6j} \otimes [X_6 + \sin(\lambda t)\alpha_6, Y_6, \cos(\lambda t)\delta\alpha_6, 1]^\top \\ l'_{6j} \otimes [X_6 + \sin(\lambda t)\alpha_6, Y_6, \cos(\lambda t)\delta\alpha_6, 1]^\top \end{bmatrix}_{12\times 12}$$

Three columns of A are multiplications of $\cos(\lambda t)$. Therefore $\cos^3(\lambda t)$ is a global scale factor of the constraint $\det(A) = 0$. Every column which contains $\sin(\lambda t)$ can be factored into a sum of the part which does not contain $\sin(\lambda t)$ and the part which contains it. This situation occurs in three columns, and by the multi-linearity of the determinant expansion we factor it, along these sums, into a sum of $2^3 = 8$ determinants. All the determinants which contain a column which is a multiplication of $\sin(\lambda t)$ vanish, since these columns are just a scalar times one of the columns which are multiplications of $\cos(\lambda t)$. Thus, the only remaining part of the determinant is the part which does not contain any $\sin(\lambda t)$ at all.

$$\det(A) = \cos^3(\lambda t) \begin{vmatrix} l_1 \otimes [X_1, Y_1, \delta\alpha_1, 1]^\top \\ l'_{1j} \otimes [X_1, Y_1, \delta\alpha_1, 1]^\top \\ . \\ . \\ l_{6j} \otimes [X_6, Y_6, \delta\alpha_6, 1]^\top \\ l'_{6j} \otimes [X_6, Y_6, \delta\alpha_6, 1]^\top \end{vmatrix}$$

The resulting constraint is a multilinear expression in the measurements p_5, p_6 and is *invariant* to time.

4 Experiments

We have applied the view-invariant polynomials and tensors for matching against classes of objects for which both shape and motion form an integral part of their characterization as a class. For example, the distinction between a person making a walking movement or a sitting movement from an upright position requires both shape (3D position of control points) and motion. Likewise, the distinction between various classes of hand gestures also requires both the 3D position of control points and their motion. In both cases, we made the assumption the dynamic component can be approximated by the non-rigidity assumed in this paper: for the gait and sitting movement we assumed multiply moving points along straight-line paths with constant velocity and for the hand gestures we assumed constraint circular trajectories with constant angular velocity.

Although our objective in this work is not necessarily to introduce the best algorithm for human gait recognition, or identification of action in general — these applications are introduced here only as a means to form interesting and challenging testing platforms — it is worth noting that the literature on understanding action — mostly related to human motion understanding covering full body movements like gait recognition up to facial expressions — is vast in number and spans a variety of different technical approaches. An updated survey of the various techniques can be found in [1, 11].

4.1 Indexing Into a “Sitting” Motion

We applied the polynomials described for the Affine camera model for the task of indexing into the motions of a person performing a sitting movement. Recall that a single view of 5 points gives rise to two view-invariant and motion-invariant polynomials whose coefficients (a function of the positions P_i and velocities V_i) can be recovered from 8 views of of the object in motion. Consider applying this scheme to the matching points arising from a person whose motion starts from an upright position and ends in a crouching position (sitting on a chair, see Fig. 1(a.1-a.3)).

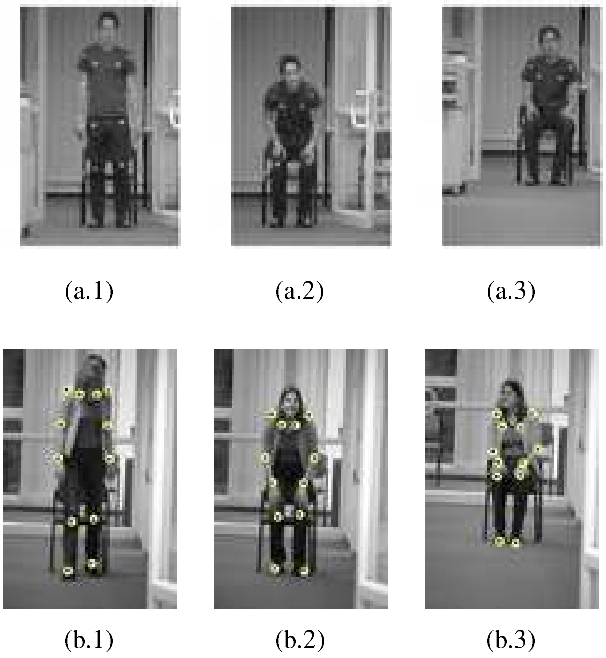


Figure 1: (a.1-a.3) shows three images from a training set of 20. The markers on the person were tracked automatically. (b.1-b.3) matched images for a novel sequence corresponding to the same "time" stamp j calculated from the image positions of the markers.

The first experiment is to recover the time j per frame of the sequence. Recall that once we have recovered the coefficients (using at least 8 views) every collection of 5 points from the image provides two 3'rd order polynomials in j, thus j can be recovered per image (for this particular setup the polynomial is linear because the velocities are mostly collinear). We consider a new sequence of another person performing the same movement and apply the polynomials from the reference person and calculate j. We then match the images from the first sequence to the images from the second sequence based on the computed j in order to see whether we indeed get the same stage in the motion sequence. Fig. 1(b.1-b.3) shows three such frames with estimated times j which match those of (a.1-a.3). We can see that there is a good correspondence between the different stages of the sitting action in the three pairs of images.

In the second experiment we display the computed j as it changes over the sequence — we expect it to change monotonically (since j represents time). Given sequence of another person performing the same type of motion we expect that the invariant polynomials recovered from the first sequence will generate a monotonically increasing j. This is shown in Fig. 2a where the value of the recovered j of the second sequence is plotted and is indeed monotonically increasing. In Fig. 2b we see the plot of j of a person getting up from a sitting position — in this case we expect to have a monotonically decreasing graph. Finally, we expect that the recovered j from a sequence of a *different* movement, say a gait movement, would yield an inconsistent behavior of j. This is shown in Fig. 2c for a walking movement — the recovered j remains flat reflecting the value (zero) corresponding to the upright position in the sequence of a person making a sitting movement. Therefore, by the change pattern of the value j one can match sequences of a class of objects making a class of movements.

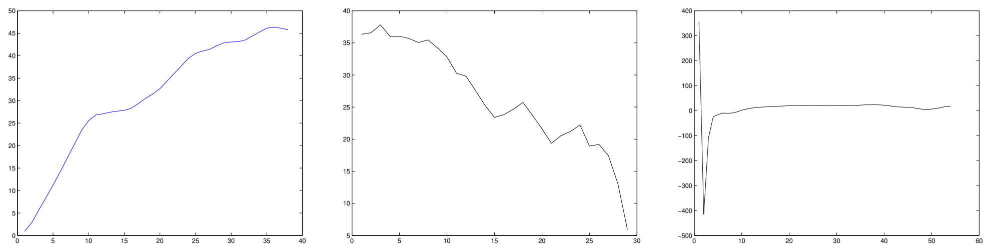
4.2 Dynamic Hand Gesture Recognition

We have created five action models — each action model was created from a sequence. The first and last frame from each sequence is displayed in the first two rows of Fig. 3. Note that some of the actions involve only rigid motion (like the hand waving in display b1 and b2) while others included dynamic (non-rigid) movement (like the gesture in displays a1 and a2). Each sequence of hand movement was taken while the camera was in motion thus we had both change of viewing position and change in shape (for those gestures that involved shape change). The stage of model building consisted of creating the indexing polynomials (the tensors) described in Section 3. The results of single view classification are presented in the graph plots. The x axis of each plot runs over the test images, whereas the y axis represents the classification values (residual of tensor contraction).

Fig. 3(f) shows the residuals for the motion shown in figures 3(b.1,b.2). The five graphs represent the residuals of the five indexing tensors. The indexing tensor matching the motion captured is emphasized for clarity. Noticeably, the indexing tensor of the correct action has much smaller residuals for most of the sequence. Fig. 3(g) shows recognition performed on a sequence capturing the motion in Fig. 3(c1,c2). In some part of this sequence the indexing function of the motion described in Fig. 3(d1,d2) show lower residuals then the correct indexing function. The images corresponding to these results are indeed compatible with both of the gestures. Fig. 3(h) shows the residuals for a sequence of the motion showed in Fig. 3(e1,e2). Two of the indexing functions (matching the motions (a) and (e)) have the lowest residuals errors, This can be explained by the fact that the motion (e) is a specific case of the motion (a), therefore a classification is not possible.

5 Summary

We have derived single-view (view-invariant) shape constraints for continuously changing 3D shapes where the notion of "time" is integrated into the shape constraints. The resulting expressions are a function of image coordinates, 3D coordinates of the corresponding points, the velocity



(a) (b) (c)

Figure 2: (a) shows how the estimate "time" j (y-axis) changes throughout a sequence (x-axis) of a sitting motion. (b) shows the same plot for a sequences of a person getting up. (c) the change in time for a sequence of a different class, in this case depicting a person walking. Note that j remains constant (around zero) which corresponds to the upright position in the sitting motion class.

vectors (assuming that the shape evolution over time follows a constant velocity rule), and the frame number j. We have shown that given a sufficient number of images, taken from various viewing positions, of the evolving shape — in the affine camera model 8 views are necessary — the coefficients of these expressions can be linearly recovered from the image measurements alone. Given a novel view of the evolving shape, the invariant expression can be evaluated leaving a polynomial in the frame number j (the time stamp) which can then be recovered. We have introduced these expressions for various camera models including 3D affine and full projective and various constant velocity evolutions models.

The experiments were conducted on sequences depicting people making a "sitting" movement and walking (gait) movements and hand gestures. In all these cases, the classes of objects are characterized both by their shape and their movement — therefore a view-invariant indexing model which takes into account shape and motion seems to be very useful. The experimental results have shown that indeed what can match various stages of the motion evolution across sequences of objects of the same class (Fig. 1), and moreover, one can efficiently discriminate between sequences coming from different classes (Fig. 2), and also make a discrimination from a single view to which class of objects the image comes from (Fig. 3).

References

[1] J.K. Aggarwal and Q. Cai Human Motion Analysis: A Review. CVIU(73), No. 3, March 1999, pp.428-440.

[2] S. Avidan and A. Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):348–357, 2000.

[3] B.M. Bennet, D.D. Hoffman, J.E. Nicola, and C. Prakash. Structure from two orthographic views of rigid motion. *Journal of the Optical Society of America*, 6:1052–1069, 1989.

[4] B.M. Bennett, D.D. Hoffman, and C. Prakash. Recognition Polynomials. *Journal of the Optical Society of America*, 10:759–764, 1993.

[5] J.W. Davis and A.F. Bobick. The Representation and Recognition of Human Movement Using Temporal Templates In *Computer Vision and Pattern Recognition (CVPR)*, June, 1997, Puerto Rico.

[6] S. Carlsson. Duality of reconstruction and positioning from projective views. In *Workshop on Representations of Visual Scenes*, 1995.

[7] S. Carlsson. Recognizing Walking People. In *Proc. of the European Conference on Computer Vision (ECCV)*, June 2000, Dublin, Ireland.

[8] S. Carlsson and D. Weinshall. Dual Computation of Projective Shape and Camera Positions from Multiple Images. *Int. J. Computer Vision* 27(3), 1998.

[9] D.T. Clemens and D.W. Jacobs. Space and time bounds on indexing 3-D models from 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:1007–1017, 1991. Also in DARPA IU Workshop, pp. 604–613, 1990.

[10] O.D. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between N images. In *Proceedings of the International Conference on Computer Vision*, Cambridge, MA, June 1995.

[11] D.M. Gavrila The Visual Analysis of Human Movement: A Survey CVIU(73), No. 1, January 1999, pp. 82-98.

[12] M. Han and T. Kanade. Reconstruction of a Scene with Multiple Linearly Moving Objects. In *Proc. of Computer Vision and Pattern Recognition*, June, 2000.

[13] A. Heyden. Reconstruction from image sequences by means of relative depths. In *Proceedings of the International Conference on Computer Vision*, pages 1058–1063, Cambridge, MA, June 1995.

[14] M. Irani and P. Anandan. Parallax geometry of points for 3d scene analyzing. In *European Conference on Computer Vision,* Cambridge, UK, April 1996.

[15] M. Irani P. Anandan and D .Weinshall. From Reference Frames to Reference Planes: Multi-View Parallax Geometry

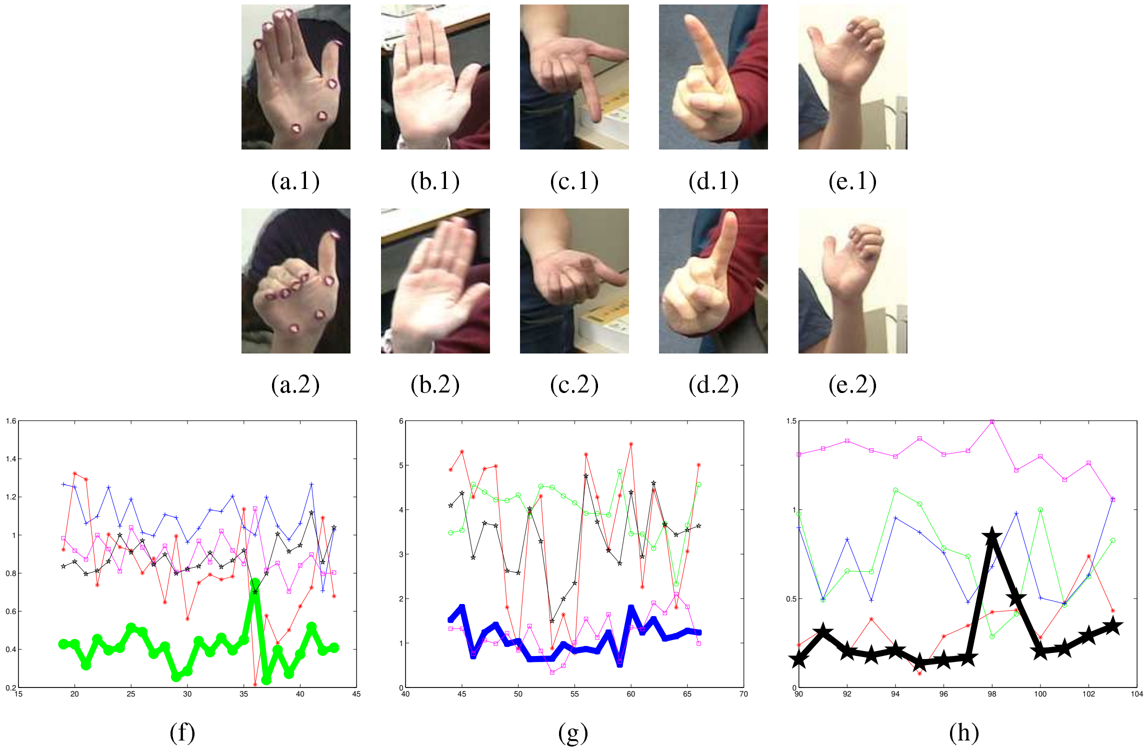


Figure 3: Dynamic hand gesture recognition from a single view using the constrained elliptical motion model. The first two rows show the first and last image of the sequence for each of the 5 classes. Some of the classes contain only rigid motion (like b.1-b.2) and the rest contain shape changes. (f,g,h) plot the residual (contraction of the indexing tensor with image measurements) of the 5 indexing tensors against frame number of the sequences. The indexing tensor of the reference class is displayed with a double-stroke. Note that in all cases the double-stroked curve has the lowest residual (modulo exceptions which are explained in the text).

and Applications. In *Proceedings of the Fifth Europen Conference of Computer Vision*, Freiburg, June 1998, Springer-Verlag.

[16] R.A. Manning and C.R. Dyer. Interpolating view and scene motion by dynamic view morphing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 388–394, Fort Collins, Co., June 1999.

[17] L. Quan. Invariants of 6 points from 3 uncalibrated images. In *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden, May 1994. Springer-Verlag, LNCS 801.

[18] A. Shashua and M. Werman. Trilinearity of three perspective views and its associated tensor. In *Proceedings of the International Conference on Computer Vision*, June 1995.

[19] A. Shashua and L. Wolf. On the Structure and Properties of the Quadrifocal Tensor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Dublin, Ireland, June 2000.

[20] A. Shashua and Lior Wolf. Homography tensors: On algebraic entities that represent three views of static or moving planar points. In *Proceedings of the European Conference on Computer Vision*, Dublin, Ireland, June 2000.

[21] B. Triggs. Matching constraints and the joint image. In *Proceedings of the International Conference on Computer Vision*, pages 338–343, Cambridge, MA, June 1995.

[22] D. Weinshall. Model based invariants for 3-D vision. *International Journal of Computer Vision*, 10(1):27–42, 1993.

[23] D. Weinshall, M. Werman, and A. Shashua. Duality of multi-point and multi-frame geometry: Fundamental shape matrices and tensors. In *Proceedings of the European Conference on Computer Vision*, LNCS 1065, pages 217–227, Cambridge, UK, April 1996. Springer-Verlag.

[24] Y. Wexler and A.Shashua. On the synthesis of dynamic scenes from reference views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, South Carolina, June 2000.

[25] L. Wolf and A. Shashua. On Projection Matrices $P^k - > P^2, k = 3, ..., 6$, and their Applications in Computer Vision. *International Conference on Computer Vision (ICCV)* Vancouver, Canada, July, 2001 .