

# View Invariants for Human Action Recognition

Vasu Parameswaran

Rama Chellappa

Center for Automation Research

University of Maryland

College Park, MD 20742

## Abstract

*This paper presents two approaches for the representation and recognition of human action in video, aiming for viewpoint invariance. The paper first presents new results using a 2D approach presented earlier. Inherent limitations of the 2D approach are discussed and a new 3D approach that builds on recent work on 3D model-based invariants, is presented. Each action is represented as a unique curve in a 3D invariance-space, surrounded by an acceptance volume ('action-volume'). Given a video sequence, 2D quantities from the image are calculated and matched against candidate action volumes in a probabilistic framework. The theory is presented followed by results on arbitrary projections of motion-capture data which demonstrate a high degree of tolerance to viewpoint change.*

## 1. Introduction

Human action analysis and recognition, has been a fertile topic of study for several years now and there exists a vast body of literature on the subject. The problem, simply stated, is to be able to accurately classify the action being performed by a human, given his/her video. Several surveys have attempted to classify various approaches for solving the problem, recent ones being [8], [7], and [4]. Though the central problem is stated quite simply, it is in fact composed of several challenging sub-problems, each of which is the subject of intense research. Broadly, these sub-problem areas are *Low level pre-processing*, *Body and pose representation* and *Action representation and recognition*. These sub-problem areas are not completely orthogonal to each other and previous approaches to the problem have not all followed such a clear-cut problem decomposition - 'holistic' approaches mapping low-level features directly to an action (e.g. [11], [5], [6]) have also been proposed. These methods may well be the best ones to use under fixed conditions but they are not applicable generally, especially for varying viewpoints.

Relative to the amount of work that has been done on human action recognition, the amount tackling viewpoint invariance has been rather small. Seitz and Dyer in [15]

have described an approach to detect cyclic motion that is affine invariant assuming that feature correspondence between successive frames is known. Rao and Shah [12] primarily target affine invariance assuming that the 2D positions of the hand are known. Syeda-Mahmood et. al. [16] represent actions as 'cylinders', formulating the problem as a joint action-recognition/fundamental-matrix recovery problem, assuming that action start and end positions are known. Rosales and Sclaroff describe an interesting approach for mapping low level features to a 2D body pose using machine learning techniques in [13]. Their approach does not require detecting or tracking body parts and results in a moderate amount of viewpoint invariance due to the inclusion of data sampled from a camera at a fixed distance all around the subject.

The manner in which body-poses and actions are represented determines the extent of applicability of an approach. Ideally, the representations should be invariant to speed of the action, frame rate of the video sequence, minor variations when performed by the same subject, minor variations when performed by different subjects and, to variations in the viewpoint. At the same time, the representation should be able to encode sufficient distinction among the various actions that we would like to be able to classify. Our work aims at achieving these objectives and explores the use of 2D and 3D invariant theory for the purpose.

It should be noted at the outset, that shot detection and other image processing tasks such as body and body-joint detection are outside the scope of the present work. Our focus is to find out how best to use the results from low-level image-processing to aid in action representation and recognition. We assume that we are given Johansson type data as input. This is achieved by the use of arbitrary projections of publicly available human motion capture data. Note that our use of motion capture data for visual human motion analysis is very different from previous approaches (e.g. [3]) in that only 2D data is used by our recognition algorithm. In other words, we use motion-capture to 'simulate' the output of a reliable low-level image-processing module.

The remainder of the paper is organized as follows: In sections 2 and 3, we briefly review a 2D method [10] and report new results. In sections 4 and 5, we discuss a more

general 3D approach which overcomes a fundamental limitation of the 2D approach and which builds on work in the area of model-based invariants [17]. We conclude and discuss future work in section 6.

## 2. A 2D Approach

It is known that there are no general, non-trivial invariants for 3D-to-2D projection. However, the theory underlying 2D-to-2D projection is very mature and there are a number of approaches one could use, if the problem could be formulated in a 2D-to-2D framework. In other words, if a 3D scene can be suitably decomposed into a number of planar patches, one could employ any convenient 2D invariant representation for the individual planes and combine the solutions. In the case of several human actions such as walking, running, waving etc., it is not difficult to identify several key poses in the action where several body joints align themselves in a plane. As an example, during walking, many joints such as the shoulders, head, feet etc. fall approximately on a plane thrice every walk cycle. Further, the left limbs and the right limbs trace areas approximately on planes during any walk cycle, provided the subject doesn't drastically alter the walking direction in that particular cycle. This is also true for running. During the sit-down action, between the instants the subject begins sitting down and that when s/he is seated, the right and left sides of the body trace planar areas. Similar arguments could be made for a number of other actions.

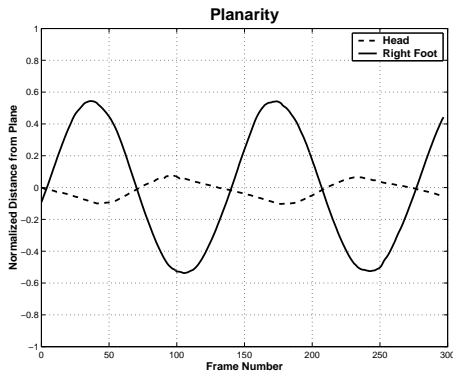


Figure 1: Plane formed by Right-Shoulder, Left-Shoulder, Left-Foot during Walking

One of the components in our action model is a *canonical pose*, where a selected set of body joints is nearly-planar. Figure 1 shows the distances (normalized by body length) of the head and right-foot to the plane formed by the joints {right shoulder, left shoulder, left foot} for a walking subject. A simultaneous zero-crossing of the curves indicates planarity. Call the body pose, when these five joints are approximately planar for walking, as  $C_1$ . This is the phase in

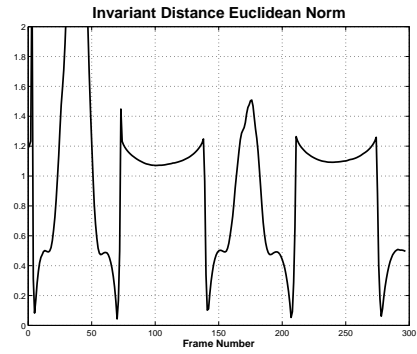


Figure 2: Euclidean Norm of Invariant Distances for  $C_1$

the action where the feet are next to each other. Given a set of five points on a plane, no three of which are collinear, there exist two invariants:

$$I_1 = \frac{M_{421}M_{532}}{M_{432}M_{521}} \quad I_2 = \frac{M_{421}M_{531}}{M_{431}M_{521}}$$

where  $M_{ijk} = |\vec{x}_i \ \vec{x}_j \ \vec{x}_k|$  and  $x_j$  is the 2D point in homogenous coordinates. For as many canonical poses as we choose to model an action with, we can pick five joints and pre-compute two such invariants using known ground-truth (using motion-capture data, for example). Given a video sequence, if the five joints are detected (or estimated) we can calculate the determinant cross-ratios and compare them against those of the canonical poses. We use the method employed in [1] to calculate the 'distance' between two invariants, weighted based on their probability distributions on the plane. Figure 2 shows the Euclidean-norm of the invariant distances of the  $C_1$  pose for the same walking-sequence used in figure 1. Using small perturbation analysis, it can be shown that small deviations in  $\vec{x}_j$  result in small deviations in  $(I_1, I_2)$ . Hence, we can use suitable thresholding of the invariant-distance Euclidean-norm, followed by a local minima selection to decide if the body is potentially in a particular canonical pose.

One of the problems with this approach is that, the fact that the invariants match those of a particular canonical pose, may not necessarily imply that the body is *in* that pose. The invariants will also match if some of the body joint world positions lie anywhere along their lines of sight from the camera. In fact, there are infinite such spurious poses. There are two strategies for reducing the probability that such spurious poses will be detected. First, we can represent a canonical pose by more than one five-tuple of joints with the observation that the probability that two sets of five joints are simultaneously in spurious position is low. Secondly, we can exploit planarity of joints spanning several frames - we already noted that certain limbs trace areas on a plane during several actions. As an example, for walking, alternate occurrences of the  $C_1$  pose delineate a sequence

of frames during which, the left-arm/left-leg and the right-arm/right-leg are on the same respective planes (approximately). Two joints for the start and end frames give us four points. The two invariants formed by a fifth moving point

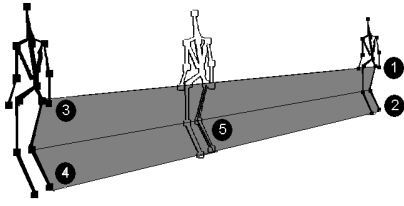


Figure 3: ISTs

will trace two trajectories which we call invariance space trajectories (ISTs) (see figure 3). Running can be handled in the same way. For the sit-down action, the delineator poses include the pose when the feet become stationary (stationarity being view invariant) in preparation for sitting down and the pose when the entire body becomes stationary. The left-leg/left-shoulder and right-leg/right-shoulder trace areas on a plane. In addition to viewpoint invariance, the use of ISTs buys us independence from the frame rate and speed of the action as well, because the starting and ending instants (and hence, duration) of the ISTs are not fixed upfront. Rather, they are event driven (i.e. determined by the occurrence of specific canonical poses). An action model would thus consist of one or more canonical pose specifications (i.e. the joint names and the two expected invariant values), and one or more ISTs between them. The complete algorithm can be found in [10].

### 3. 2D Results

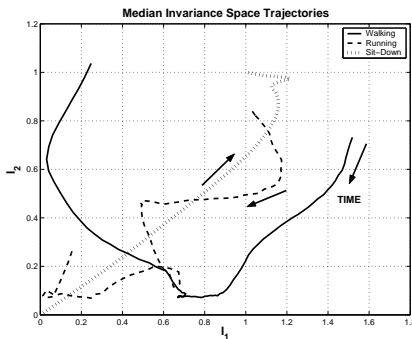


Figure 4: ISTs for walking, running and sitting-down

We obtained motion-capture data from Credo Interactive Inc. and Carnegie Mellon University in the BioVision Hierarchy and Acclaim formats. The combined dataset included

12 subjects performing 25 walking sequences, 6 running sequences and 18 sit-down sequences. Each walking and running sequence included one to three complete cycles of the respective action. There is very little variability in the invariant values for  $C_1$  across subjects. We also found that the  $C_1$  invariants are very similar between walking and running and hence are by themselves, not sufficient to distinguish walking from running. However, the ISTs are markedly different. For walking and running, the same set of joints were used - the hip and foot were chosen as the two joints at delineator frames while the knee was chosen as the moving point. For the sit-down action, the head and foot were chosen at the delineator frames while the hip was chosen as the moving point. Figure 4 shows the ISTs for the three actions. Five random viewpoints were chosen for evaluation

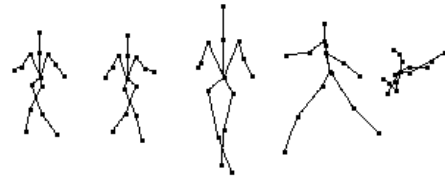


Figure 5: Walk seq. 20, frame no. 100, all viewpoints

of the algorithm. Figure 5 shows the same walking frame as seen from the five different viewpoints. The action-model (i.e. the  $C_1$  invariants and the ISTs) were calculated in a fully automatic manner from the available motion-capture data. A planarity assessment program was first run on the dataset to determine those frames at which the joints for  $C_1$  were closest to a plane. There are two possible cases when the body is in  $C_1$  - either the left or right leg can be stationary while the other leg is in motion. A convention was adopted that the  $C_1$  pose where the left leg is in motion is the ‘start’ of the action (for both, walking and running). This also enabled us to automatically determine the ‘ground-truth’. ISTs were automatically calculated between two consecutive starting  $C_1$  poses. The algorithm was run on all action-instances (792 in number). The distance of each action-instance was computed against each of the three action-models, and the action-model with the least distance was declared to be the classified action. This was evaluated against the known ground-truth. The following table summarizes the action classification results:

Metric	Viewpoint Number				
	1	2	3	4	5
Tot. Detects	795	780	703	719	646
True Detects	764	757	689	708	640
True Det. %	96.46	95.58	86.99	89.39	80.80
Misclass.	31	23	14	11	6
Misclass. %	3.89	2.94	1.99	1.53	0.92

The true detection rate here is defined as the ratio of correct detections to the expected number of detections (792 in this case). A misclassification is defined as an instance where the action was classified incorrectly (e.g. a walk-cycle was classified as a run-cycle or a sit-down action). The misclassification rate is defined as the ratio of the number of misclassifications to the number of total detections. As it can be seen, the performance of the algorithm for viewpoint 5 is the worst among all the viewpoints. The viewpoint is such that the camera looks down on the subject, and there are several instances where some of the body joints are coincident or very close to each other which amplifies the errors. In many cases, the  $C_1$  pose was missed and this contributed to the poor results.

## 4. A 3D Approach

Although many actions can be decomposed into planes for a suitable solution by way of the 2D approach, the approach itself is not sufficiently general for the modeling of a general action, necessitating the use of a 3D approach. In [2] the authors show that there are no general, non-trivial 3D invariants that can be used to recognize an arbitrary 3D point-set. However, this does not imply that invariants cannot be used for 3D object recognition. Indeed, as was shown by several researchers, notably by Rothwell et al [14] and later by Weiss and Ray [17] it is possible to recognize 3D objects from a single view. Weiss and Ray in [17] developed relationships between six-tuple 3D points and their corresponding image coordinates that are satisfied for all views of the 3D points. We first their main result below:

Given a set of six world points  $\{\vec{X}_i\}$ , atleast the first four of which are not on a plane, and their corresponding image points  $\{\vec{x}_i\}$ , (both sets of points in homogenous coordinates) the following relation holds:

$$I_3 (I_2 - 1) i_1 i_2 - I_3 (I_1 - 1) i_1 - I_1 (I_2 - 1) i_2 - I_2 (I_3 - 1) i_3 i_4 + I_2 (I_1 - 1) i_3 + I_1 (I_3 - 1) i_4 = 0 \quad (1)$$

where

$$i_1 = \frac{m'_{12} m_{14}}{m_{12} m'_{14}} \quad i_2 = \frac{m'_{12} m_{35}}{m'_{13} m_{25}} \quad i_3 = \frac{m'_{12} m_{13}}{m_{12} m'_{13}} \quad i_4 = \frac{m'_{12} m_{45}}{m'_{14} m_{25}} \\ I_1 = \frac{M_1 M'_2}{M'_1 M_2} \quad I_2 = \frac{M_1 M'_3}{M'_1 M_3} \quad I_3 = \frac{M_1 M'_4}{M'_1 M_4} \quad (2)$$

Here  $M_1 = |\vec{X}_2 \vec{X}_3 \vec{X}_4 \vec{X}_5|$  i.e. the determinant formed by the first five points after removing the point of the index. The prime denotes the same quantity with the sixth point substituted for the fifth point, i.e.  $M'_1 = |\vec{X}_2 \vec{X}_3 \vec{X}_4 \vec{X}_6|$ .  $M_i$  determine the model invariants.  $m_{ij}$  determine the image plane determinants. Indexing similarly by the ‘points

left out’ we use  $m_{12}$  for  $|\vec{x}_3 \vec{x}_4 \vec{x}_5|$  and  $m'_{12}$  for  $|\vec{x}_3 \vec{x}_4 \vec{x}_6|$ .  $m_{ij}$  and  $m'_{ij}$  are quantities derived from image coordinates. Hence, equation 1 describes a compatibility relation between the six-tuple 3D points and their 2D image points, that holds for all viewpoints.

The human body can be represented by as many six-tuples as possible, corresponding to several body joints. Each such six-tuple will give rise to three 3D invariant values. For the human body performing an action, the invariant values will vary temporally as the action evolves, and give rise to a 3D curve in invariance-space for each six-tuple. When the same action is performed by different subjects, the invariance space curves so traced will be slightly different from each other. Furthermore, the same action performed by the same subject multiple times will also result in slightly different curves, as humans are not perfectly consistent in performing actions. Hence, the action will have to be represented not only as a 3D curve through invariance space, but also with a surrounding ‘acceptance volume’ around the curve based on the probability density functions (pdfs) of the invariants. Let’s term the representation as an ‘action-volume’. Let us also term the 3D curve as an ‘action-curve’. Using ground-truth from motion capture, action-volumes can be empirically estimated for several actions, resulting in an action database. The 3D curve itself can be defined as the median curve obtained from training data. Parametrically, the 3D curve can be represented by three 2D curves with the abscissa being non-dimensionalized time in  $[0, 1]$ , and the ordinate being the invariants.  $t = 0$  will represent the start while  $t = 1$  will represent the end of the action.  $t$  represents the phase of the action. The 2D curves can be represented by cubic splines for  $N - 1$  intervals. For the  $k$ th such interval:

$$I_j(t) = a_{jk0} + a_{jk1}t + a_{jk2}t^2 + a_{jk3}t^3$$

where  $t_k \leq t \leq t_{k+1}$   $t_1 = 0, t_N = 1$ . For each knot of the spline, the empirically determined pdf  $p((I_1, I_2, I_3)|A_i, t_k)$  for a known action  $A_i$  and phase  $t_k$  is known to us.

Given a video sequence consisting of a labeled set of body joints and their image locations per frame, we can calculate the image based invariants  $i_j$  in each frame. When substituted in (1), these 2D invariants will determine a quadric surface in 3D invariance space whose variables are the (unknown) 3D invariants:

$$\alpha^T \mathbf{A} \alpha + \alpha^T \mathbf{b} = 0 \quad (3)$$

where

$$\alpha = [I_1 \ I_2 \ I_3]^T \\ \mathbf{b} = [i_2 - i_4, \ i_3 i_4 - i_3, \ -i_1 i_2 + i_1]^T \\ \mathbf{A} = \begin{bmatrix} 0 & (-i_2 + i_3)/2 & (-i_1 + i_4)/2 \\ (-i_2 + i_3)/2 & 0 & (i_1 i_2 - i_3 i_4)/2 \\ (-i_1 + i_4)/2 & (i_1 i_2 - i_3 i_4)/2 & 0 \end{bmatrix}$$

The quadric surface will potentially intersect several action volumes and in particular, the action-volume corresponding to the action being performed. Equation 3 essentially constrains the set of possible actions that the body is performing, to those that the quadric surface intersects. Each intersection point is a candidate phase of a candidate action. A problem to be solved here is that of finding the action-volumes in the database that the quadric surface intersects. Allowing for variabilities, we also need a measure of closeness of the quadric surface to each phase of each candidate action-volume. Ignoring probabilities for now, if the three time-parametrized cubic splines for each action-curve are substituted into 3, the problem becomes that of finding zeroes of a sixth degree polynomial in  $t$  per interval and checking if the roots fall within the interval. This would have to be repeated for all intervals of the cubic spline.

The problem can be simplified if we sacrifice perfect-accuracy and settle for linear-splines. In addition, computational cost can be reduced by operating on a coarser discretization of the time interval  $[0, 1]$ . This would result in a quadratic equation in  $t$ . If the quadric doesn't intersect the given action volume, the quadratics corresponding to the action will have imaginary roots. In this case, we would need to determine the point on the quadric surface that is closest to the line segment representing the interval. We seek the cylinder of smallest radius with axis as the line segment that is tangent to the quadric. To solve this problem, we first transform the quadric and the action curve to a canonical frame of reference where the line segment runs from  $(0, 0, 0)$  to  $(1, 0, 0)$  using translation, rotation and scaling. Note that these are Euclidean transformations of the 3D invariance space. The matrix form of the quadric allows us to easily calculate its equation in the canonical coordinate system by substituting  $\alpha = \mathbf{sR}(\bar{\alpha} - \mathbf{t})$  into equation 3 giving:

$$\bar{\alpha}^T \bar{\mathbf{A}} \bar{\alpha} + \bar{\alpha}^T \bar{\mathbf{b}} + \bar{c} = 0 \quad (4)$$

The problem becomes one of minimizing  $\bar{I}_2^2 + \bar{I}_3^2$  subject to  $0 \leq \bar{I}_1 \leq 1$  and (4). Lagrange multipliers and slack variables for the inequality constraints can be used to arrive at a solution to the problem. This process involves the solution of a quartic equation which in turn requires the solution of a cubic equation, an elegant solution to which can be found in [9]. The solution is transformed back into the original frame of reference and the optimal point on the quadric  $(I_1^*, I_2^*, I_3^*)$  is found. The optimal phase  $t^*$  is that which produces  $I_1^*$ . We look up the known joint probability  $p((I_1^*, I_2^*, I_3^*)|A, t^*)$  where  $A$  is the action. At each video frame, for all the phase intervals, we thus calculate the probability that a particular pose of a particular action is occurring at that frame. As an example, figure 6 shows the probability for the starting pose ( $t^* = 0$ ) of the running action as seen from viewpoint 1 (from the viewpoints used in section 3), smoothed using a gaussian filter of width 7.

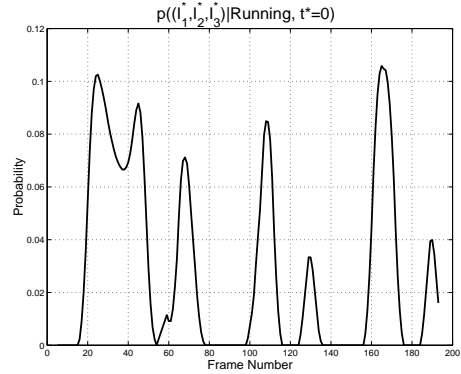


Figure 6: Probability of the starting pose for the running action

There were two difficulties we encountered with this approach. First, experiments with motion capture data revealed that the action volumes for walking and running were very similar, and not by themselves adequate for distinguishing between them, perhaps pointing to a loss of specificity in representing the action in 3D invariance space. Second, we initially pursued a dynamic programming approach where the probability of the action occurring between two given frames was maximized. For continuous action sequences, this was problematic in the sense that the start of one instance and the end of a later instance was sometimes returning a probability greater than that for the true start and end. Although this could be fixed using heuristics, we settled for a different approach described below, that also addresses the first problem.

In addition to maintaining action volumes, we maintain a Euclidean representation of the action. From a given video sequence, we extract 2D quantities and calculate the probability of the start and end phases per frame as described above. We pick frames where the probabilities are above a threshold and are local maxima. (1) establishes a correspondence between the world points and image points for six points (i.e. 12 equations), which is enough to be able to determine the world-to-image transform matrix of 11 unknowns. However, we have twelve point correspondences (i.e. 24 equations) from the starting and ending poses, which we use to estimate the matrix using Singular Value Decomposition. Given this, the Euclidean representation of the action is transformed to the image coordinate system and the differences between joint positions are calculated. If the differences are within a threshold, the action is declared as being potentially occurring between the start and end frames. If more than one actions match, the one with the smaller distance is chosen. Picking the right image threshold is important. Fixing it to a constant for all viewpoints would make the algorithm throw out perfectly good matches for one viewpoint while possibly accepting

false matches for others. One idea would be to choose an absolute threshold that is ground-truth based and *calculate* viewpoint thresholds based on the absolute threshold and the estimated world-to-image transform. We allow a 3D joint position to be within an error cube. The 3D threshold is the side of the cube, which can be fixed upfront. If necessary, one can use different 3D thresholds for different joints of the body, based upon their relative importance for the action. Let us consider one particular joint  $J$  whose expected 3D position is  $(X_0, Y_0, Z_0)$ . Let us say that we have equal thresholds  $2\epsilon$  such that we allow the 3D position of  $J$ ,  $(X, Y, Z)$  to be within the following cube:

$$|X - X_0| \leq \epsilon \quad |Y - Y_0| \leq \epsilon \quad |Z - Z_0| \leq \epsilon \quad (5)$$

Let us say that the observed 2D position is  $(x, y)$  while the expected 2D position is  $(x_0, y_0)$ . The goal is to determine a threshold for the difference,  $(\Delta x, \Delta y) = ((x - x_0), (y - y_0))$ . We further know the estimated world-to-image transform  $T$ . Let  $X = X_0 + \delta_x$ ,  $Y = Y_0 + \delta_y$ ,  $Z = Z_0 + \delta_z$ . We have the following:

$$\Delta x = \left( \frac{T_{11}(X_0 + \delta_x) + T_{12}(Y_0 + \delta_y) + T_{13}(Z_0 + \delta_z) + T_{14}}{T_{31}(X_0 + \delta_x) + T_{32}(Y_0 + \delta_y) + T_{33}(Z_0 + \delta_z) + 1} \right) - \left( \frac{T_{11}X_0 + T_{12}Y_0 + T_{13}Z_0 + T_{14}}{T_{31}X_0 + T_{32}Y_0 + T_{33}Z_0 + 1} \right) \quad (6)$$

$$\Delta y = \left( \frac{T_{21}(X_0 + \delta_x) + T_{22}(Y_0 + \delta_y) + T_{23}(Z_0 + \delta_z) + T_{24}}{T_{31}(X_0 + \delta_x) + T_{32}(Y_0 + \delta_y) + T_{33}(Z_0 + \delta_z) + 1} \right) - \left( \frac{T_{21}X_0 + T_{22}Y_0 + T_{23}Z_0 + T_{24}}{T_{31}X_0 + T_{32}Y_0 + T_{33}Z_0 + 1} \right) \quad (7)$$

Neglecting higher order terms, after simplification, we obtain:

$$\frac{\Delta x}{x} = \frac{\Delta N_x}{N_x} - \frac{\Delta D}{D} \quad (8)$$

$$\frac{\Delta y}{y} = \frac{\Delta N_y}{N_y} - \frac{\Delta D}{D} \quad (9)$$

where

$$\begin{aligned} N_x &= T_{11}X_0 + T_{12}Y_0 + T_{13}Z_0 + T_{14} \\ N_y &= T_{21}X_0 + T_{22}Y_0 + T_{23}Z_0 + T_{24} \\ D &= T_{31}X_0 + T_{32}Y_0 + T_{33}Z_0 + 1 \\ \Delta N_x &= T_{11}\delta_x + T_{12}\delta_y + T_{13}\delta_z \\ \Delta N_y &= T_{21}\delta_x + T_{22}\delta_y + T_{23}\delta_z \\ \Delta D &= T_{31}\delta_x + T_{32}\delta_y + T_{33}\delta_z \end{aligned}$$

Given that  $|\delta_x| < \epsilon$ ,  $|\delta_y| < \epsilon$ ,  $|\delta_z| < \epsilon$ , the above equations determine the acceptable values for  $(\Delta x, \Delta y)$  in terms of known or estimated quantities. Thus, we automatically achieve viewpoint based thresholding via the involvement of  $T$  and need only fix the 3D threshold  $\epsilon$  upfront. We found that this combined projective-space/Euclidean-space representation worked well at distinguishing walking from running.

The following is an outline of the algorithm. We are given an action  $A$  which consists of the action volume  $M$  and the Euclidean action model  $E$ . We are also given 2D image data in frames,  $\{F\}$ . The output is a list of start and end positions where the action is found to occur:

1. For each frame  $f \in F$ 
  - (a) Calculate  $i_1, i_2, i_3, i_4$  which determine the quadric surface (equation (3)).
  - (b) For the starting phase interval of  $M$ , calculate quadratic and find roots if any. If no roots, calculate point on quadric, closest to the phase interval by using Lagrange multipliers (section 4). In both cases, the optimal point in 3D invariance space  $I_1^*, I_2^*, I_3^*$  and the optimal phase  $t^*$  is found from which, the probability  $p((I_1^*, I_2^*, I_3^*)|A, t^*)$  is looked-up and pushed into a list  $L$  indexed by frame number.
2. Apply temporal domain smoothing to  $L$  and extract local maxima. Push extracted local minima into list  $L^*$ . Clear  $L$ .
3. For indices  $(s, e) \in L^*$  where  $s < e$ 
  - (a) Estimate world-to-image transform  $T$  of  $E$  to the image plane, with  $(s, e)$  corresponding to the start and end of  $E$  and transform all of  $E$  to the image plane.
  - (b) Calculate deviation in predicted and observed joint 2D locations and do thresholding based on equations (8) and (9). If acceptable, push  $(s, e, d)$  into list  $L$ , where  $d$  is the net deviation.
4. Remove overlapping intervals from  $L$  retaining those with smaller deviation and return  $L$ .

## 5. Results using 3D Approach

Performance evaluation was done for the walking and running actions. The same data, ground-truth and viewpoints were used as for the 2D approach. The action volumes and the Euclidean space representations of the actions were calculated automatically and input to the 3D algorithm as the action model. The median curve was used for both the representations. The same six joints used for both, walking and running and were the following : {Right-Knee, Right-Foot, Right-Elbow, Left-Knee, Left-Hand, Left-Foot}. The following table shows the results.

Metric	Viewpoint Number				
	1	2	3	4	5
Tot. Detects	528	522	510	542	556
True Detects	512	506	481	537	476
True Det. %	88.88	87.85	83.51	93.23	82.64
Misclass.	16	16	29	5	80
Misclass. %	3.03	3.07	5.69	0.92	14.39

The true detection rate and misclassification rate definitions remain the same as that for the 2D approach. The total number of ground-truthed actions is 576 in this case. As in the 2D case, performance for viewpoint 5 is the worst, although the difference in detection rate from those of other viewpoints is not as great as for the 2D case. Relative to other viewpoints and to the 2D case, false positives are significantly higher. This can perhaps be attributed primarily to the peculiarities of viewpoint itself (camera looking down, joints coincident or in very close proximity etc.). Performance for some of the other viewpoints is worse than that for the 2D approach, suggesting that for select actions for which planar decomposition is possible, the 2D approach is preferable. However, of course, the 3D approach is more applicable for a general action. Considering the variety of viewpoints chosen, the 3D results look encouraging.

## 6. Conclusions

We presented two approaches for modeling and recognizing human actions that are highly tolerant to viewpoint change, using 2D and 3D invariance theory. The approaches are also independent of action-speed/frame-rate, subjects, and minor variabilities in the action. Experimental results on 2D projections of motion capture data show good success rates. As such, the analysis and results are preliminary and warrant further investigation and performance analysis. Clearly, the action vocabulary needs to be broadened (modeling of the sit-down action for the 3D case is still in progress and many other actions need to be modeled as well). The discriminating power of action-volumes in representing actions, and the possible use of more formal event based models (e.g. Petri Nets or coupled Hidden Markov models) is being investigated. Analysis is also underway to formulate the problem as a joint tracking/action-recognition problem using 3D invariants and we believe that a state-space based approach that encodes the overall pose of the body and not specific ‘image-features’ which are sensitive to image noise, is worth investigating.

## Acknowledgments

This work was supported in part by NSF Grant ECS 02-25475.

## References

- [1] K.E Astrom and L. Morin. Random cross ratios. *Proc. 9th Scand. Conf. on Image Analysis*, June 1995.
- [2] J. B. Burns, R. S. Weiss, and Riseman E. M. The non-existence of general case invariants. In J. L. Mundy and A. Zisserman, editors, *Geometric Invariance in Machine Vision*. MIT Press, Cambridge, MA, 1992.
- [3] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. *Proc. International Conference on Computer Vision*, pages 624–630, 1995.
- [4] C. Cedras and M. Shah. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2), 1995.
- [5] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997.
- [6] H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletonization. *Fourth IEEE Workshop on Applications of Computer Vision*, pages 15–21, 1998.
- [7] D. M. Gavrilu. The Visual Analysis of Human Movement. *Computer Vision and Image Understanding*, 73(1):82–98, 1998.
- [8] T. Moeslund and E. Granum. A survey of computer vision based human motion capture. *Computer Vision and Image Understanding*, 81(3), March 2001.
- [9] R. W. D Nickalls. A new approach to solving the cubic: Cardan’s solution revealed. *The Mathematical Gazette*, 77, 1993.
- [10] V. Parameswaran and R. Chellappa. Quasi-invariants for human action representation and recognition. *Proc. International Conference on Pattern Recognition*, 2002.
- [11] R. Polana and R. C Nelson. Detecting Activities. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–7, 1993.
- [12] C. Rao and M. Shah. View-invariant representation and learning of human action. *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, July 2001.
- [13] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.
- [14] C.A. Rothwell, D.A. Forsyth, A. Zisserman, and J.L. Mundy. Extracting projective structure from single perspective views of 3d point sets. *Proc. International Conference on Computer Vision*, pages 573–582, 1993.
- [15] S. M. Seitz and C. R Dyer. View-invariant analysis of cyclic motion. *International Journal of Computer Vision*, 25:1–25, 1997.
- [16] T. Syeda-Mahmood and A. Vasilescu. Recognizing action events from multiple viewpoints. *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, July 2001.
- [17] I. Weiss and M. Ray. Model-based recognition of 3d objects from single images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23, February 2001.