On the use of Anthropometry in the Invariant Analysis of Human Actions

Alexei Gritai, Yaser Sheikh and Mubarak Shah
School of Computer Science
University of Central Florida
Orlando, FL-32826
USA
{agritsay, yaser, shah}@cs.ucf.edu

Abstract

In this paper, we propose a novel approach to matching human actions using semantic correspondence between human bodies with an eye towards invariant analysis of activity. The correspondences are used to provide geometric constraints between multiple anatomical landmarks (e.g. hands, shoulders and feet) to match actions performed from different viewpoints and in different environments. The fact that the human body has certain anthropometric proportion allows innovative use of the machinery of epipolar geometry to provide constraints to accurately analyze actions performed by different people leading to some interesting results. Temporally invariant matching is performed, using non-linear time warping, to ensure that similar actions performed at different rates are accurately matched as well. Thus, the proposed algorithm guarantees that both temporal and view invariance is maintained in matching. We demonstrate the versatility of our algorithm in a number of challenging sequences and applications.

1 Introduction

Invariants are properties of geometric configurations that remain unaffected by a certain class of transformations. In the context of action recognition, it is desirable that algorithms maintain invariance to view and also to rate of execution, or in other words, invariance with respect to the set of possible projection matrices and with respect to temporal transformation. While invariance in object recognition has an established body of work, [9], invariance in *action* recognition has received relatively little attention. Action recognition has primarily been performed using image information, such as correlation [5], trajectory matching [10], histogram intersection [13] and all these methods are dependent, to different degrees, on the view at which an action is observed. Recent work on view invariant action recognition [12] has addressed this issue by using invariant metrics. However, this algorithm represents an action by a single point, which as a means of representation is limited and ambiguous since an action is rarely sufficiently described by a single point.

In this paper, we address the invariant recognition of human actions, and investigate the use of anthropometry to provide constraints on matching. The study of human proportions has a great tradition in science, from the 'Golden Sections' of ancient China, India, Egypt and Greece down to renaissance thinkers like Leornardo Da Vinci (the Vitruvian Man) and Albrecht Durer, with modern day applications in Ergonomics and human performance engineering. We make implicit use of the 'laws' governing human body proportions to provide geometric constraints for matching. Instead of using a single point representation, we explore the use of several points on the actor for action recognition, and use geometric constraints with respect to two *actors* performing the action instead of two camera *views*. This innovative use of geometry allows two interesting results for the recognition of actions. The first result provides a constraint to measure the similarity of the posture of two actors viewed in two images. The second result extends this first constraint to globally measure similarity between two actions. These results are described in Section 2. We perform experiments on a particularly challenging set of images, with actions performed at different rates by people of different sizes, races and sex, taken from different view points. The experimental results are presented in Section 4 with a discussion of conclusions in Section 5.

2 The Analysis of Actions

In this section we discuss our representation of actions and propose a novel matching scheme based on semantic correspondences between humans. Geometric constraints on these correspondences are used to analyze actions as they occur. The main concern in the presented work is the recognition of human activity performed by different people at varying rates in different environments or viewpoints. A primary feature is that our recognition measurement is invariant to both view-point and to the speed of execution.

2.1 Representation of Actors and Actions

The model of a moving body as a point is ubiquitous in the physical sciences community. In [8], Johansson demonstrated that a simple point-based model of the human body contained sufficient information for the recognition of actions. Relying on this result, we represent the current pose and posture of an actor in terms of a set of points in 3-space in terms of a set of 4-vectors $\hat{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2 \dots \mathbf{X}_n\}$, where $\mathbf{X}_i = (X_i, Y_i, Z_i, \Lambda)^\top$ are homogenous coordinates. A *posture* is a stance that an actor has at a certain

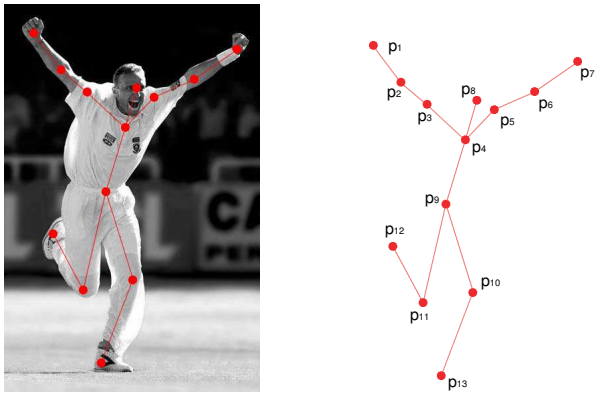

Figure 1. Point-based representation. Johansson's experiments in [8] demonstrate that point-based representations contains sufficient information for action recognition.

time instant, not to be confused with the actor's *pose*, which refers to position and orientation (in a rigid sense). Each point represents the spatial coordinate of an anatomical landmark (see [2]) on the human body as shown in Figure 1. The imaged pose and posture are represented by $\hat{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n\}$, where $\mathbf{x}_i = (x_i, y_i, \lambda)^\top$. $\hat{\mathbf{X}}$ and $\hat{\mathbf{x}}$ are related by a 4×3 projection matrix $\mathtt{C}$, i.e. $\hat{\mathbf{x}} = \mathtt{C}\hat{\mathbf{X}}$. As will be seen presently, eight imaged points on human body are required in each frame of video and, at least, one of them must correspond to the body part directly involving in action. We refer to each entity involved in an *action* as an *actor*. An *action element* is the portion of an action that is performed in the interval between two frames. Each action is represented as the set of action elements $\hat{\mathbf{U}} = \{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots \hat{\mathbf{u}}_t\}$, where $\hat{\mathbf{u}}_t = \{\mathbf{u}_1, \mathbf{u}_2, \dots \mathbf{u}_n\}$ is the set of motion vectors in the real world that define an action element between time t and $t+1$. For a comparison of other representations to this one the reader is referred to [6].

2.2 Anthropometric Constraints

Both body size and proportion vary greatly between different races and age groups and between both sexes. However, while human dimensional variability is substantial, several anthropometric studies (see [4], [3], [1]) empirically demonstrate that it is not *arbitrary*. These studies have tabulated various percentiles of the dimensions of several human anatomical landmarks. In this paper, we conjecture that for at least 90% of the human population the proportion between human body parts can be captured by a projective transformation of $\mathbb{P}^3$.

Conjecture 1 Suppose the set of points describing actor A_1 is $\hat{\mathbf{X}}$ and the set of points describing actor A_2 is $\hat{\mathbf{Y}}$. The relationship between these two sets can be described by a matrix $\mathcal{M}$ such that

$$\mathbf{X}_i = \mathcal{M}\mathbf{Y}_i \tag{1}$$

where $i = 1, 2 \dots n$ and $\mathcal{M}$ is a 4×4 non-singular matrix.

This was empirically supported using the data in [2] (Table 5-1 and 5-2 which record the body dimensions of male and female workers between the ages of 18 and 45). Between the dimensions of the '5th percentile woman' and the '95th percentile man', where a mean error of 227.37 mm was found before transformation, a mean error of 23.87 mm was found after applying an appropriate transformation. Using this property, geometric constraints can be used between the imaged points, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ of the two actors. The transformation $\mathcal{M}$ simultaneously captures the different pose of each actor (with respect to a world coordinate frame) as well as the difference in size/proportions of the two actors.

2.2.1 Postural Constraint

If two actors are performing the same action, the postures of each actor at a corresponding time instant (with respect to the action time coordinate) should be the same. Thus an action can be recognized by measuring the similarity of posture at each corresponding time instant.

Proposition 1 If $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{y}}_t$ describe the imaged posture of two actors at time t, a fundamental matrix $\mathcal{F}$ can be uniquely associated with $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ if the two actors are in the same posture.

It is known (pg. 247 Section 9.2, [7]) that for uncalibrated cameras the ambiguity of structure is expressed by such an arbitrary non-singular projective matrix. If two actors are in the same posture, the only difference between their point-sets is a projective relationship (Conjecture 1). Thus, if an invertible matrix $\mathcal{M}$ exists between $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\mathbf{X} = \mathcal{M}\mathbf{Y}$, a fundamental matrix is uniquely determined by $\mathbf{x}^\top \mathcal{F} \mathbf{y} = 0$ (Theorem 9.1 [7]).[1]

Since the labels of each point are assumed known, *semantic* correspondence (i.e. the left shoulder of A_1 corresponds to the left shoulder of A_2) between the set of points is also known. Proposition 1 states that the fundamental matrix computed using these semantic correspondences between actors inherently captures the difference in anthropometric dimensions and the difference in pose. However, in order to use this constraint in action matching, it is necessary to have a similarity metric. The similarity metric, in particular, measures the similarity between the *postures* of the two actors.

$$\mathcal{A}\mathbf{f} = \begin{bmatrix} x'_1x_1 & \dots & x'_nx_n \\ x'_1y_1 & \dots & x'_ny_n \\ x'_1 & \dots & x'_n \\ y'_1x_1 & \dots & y'_nx_n \\ y'_1y_1 & \dots & y'_ny_n \\ y'_1 & \dots & y'_n \\ x_1 & \dots & x_n \\ y_1 & \dots & y_n \\ 1 & \dots & 1 \end{bmatrix}^\top \mathbf{f} = 0 \tag{2}$$

where $\mathbf{f}$ is a 9-vector (in row-major order) representation of $\mathcal{F}$. Since Equation 2 is a homogenous equation, $\mathcal{A}$ has a rank of at most eight, if the two actors are indeed in the same posture. The similarity of the postures of two actors can then be measured using the ninth singular value of $\mathcal{A}$[2].

[1] Points that lie on the line joining the principal points are excluded.

[2] There exist some configurations for which the rank of $\mathcal{A}$ may be less than eight, but these form special cases and they can usually be ignored.

Figure 2. Frames corresponding to 'Picking Up' in six sequences. The top left frame corresponds to the example sequence, the rest are the tested sequences. In each sequence, the actors are in markedly different orientations with respect to the camera.

2.2.2 Action Constraint

Along with the frame-wise measurement of postural similarity, it is observed here that a strong *global* constraint can be used on the point sets describing two actors if they are performing the same action.

Proposition 2 For an action-element $\hat{\mathbf{u}}_t$, the fundamental matrices associated with $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ and $(\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{y}}_{t+1})$ are the same if both actors perform the action element defined by $\hat{\mathbf{u}}_t$.

Based on Conjecture 1, we can say that $\mathcal{M}$ remains the same between time t and $t+1$. In other words, $\mathcal{M}$ determines $\mathbf{Y}$ with respect to $\mathbf{X}$ and does not depend on the motion of $\mathbf{X}$. Since $\mathcal{M}$ is the same then the fundamental matrices, $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$, corresponding to $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ and $(\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{y}}_{t+1})$ are the same (p.235 Result 8.8, [7]).

Essentially, what this means is that if both individuals perform the same action-element between frame f_t and frame f_{t+1}, the transformation that captured the difference in pose and dimension between the two actors remains the *same*. As a direct consequence, the subspace spanned by the measurement matrix $\mathcal{A}$ also remains the same and this suggests that if a measurement matrix were constructed using *all* the corresponding points over the entire action $\hat{\mathcal{A}} = [\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_k]$, ideally, the ninth singular value of $\hat{\mathcal{A}}$ should be still be zero and can be used as a global measure of *action* similarity.

3 Temporally Invariant Matching

While view invariance is required in action analysis due to the physics of the imaging process, temporal invariance, i.e. invariance of action analysis to temporal transformations, is needed due to the nominal uniqueness of each actor's execution of an action. Rather than use features that are invariant to a class of temporal transformation, we loosely use the term temporally invariant matching to mean matching that is invariant to a class of temporal transformations. In this work, we perform matching invariant to a class of nonlinear time warps using Dynamic Time Warping. Dynamic Time Warping (DTW) has previously been used in [12] for video alignment, and for an introduction to the use of DTW the reader is directed to [11]. Dynamic Time Warping is particularly suited to action recognition, since it is expected that different actors may perform some portions of an action at different rates, relatively. As is demonstrated in Section 4, matching using Dynamic Time Warping is highly effective in compensating for this variability.

For certain applications, particularly when the pattern that is to be recognized is of a short duration, DTW does not provide significant improvements over a linear model of temporal transformation. It was found that the use of a linear model is also appropriate for *coarse* matching and synchronization. In the presented work, the use of DTW is not trivial since both the local (postural) constraint and the global (action) constraint need to be incorporated in computation of the similarity measure. Applying a temporal window (k frames before and after the current one) for computation of similarity measure between two agents provided a marked improvement.

4 Experimental Results

To validate the proposed work, we performed experiments in several challenging scenarios. Due to the limitation of space, the results of only three experiments have been included here. The first set of experiments involved action detection in a long sequence, the second set involved synchronizing videos to match actions, and the third set applied the proposed approach to gait analysis.

4.1 Action Detection

In this experiment, actors performed a sequence of actions: walking, bending down to grasp an object, lifting the object and walking away. Videos were taken of three different people from three different views, and the action of picking up an object was detected in each video by matching a shorter pattern sequence. Figure 2(a) shows the corresponding frames in six videos. The sensitivity of the detection was also tested in a sequence containing four individuals walking. Sample frames from this sequence are shown in Figure 2(b). A test pattern of a single cycle of the distinctive 'Egyptian' gait was compared to each actor's motion and the variation of the ninth singular value over time for each of the four actors is shown in Figure 4. There are two interesting features that can be observed in this figure. Firstly, since the posture involved in the 'Egyptian' gait is relatively distinct from the usual human gait the ninth singular value for the third actor is consistently lower than the other actors. Secondly, the sinusoidal nature of the plot clearly shows the periodicity that is associated with walking.

4.2 Action Synchronization

Three actors jumped asynchronously in the field of view of a stationary camera. The objective in this experiment was to align the actors jumps and twists so that a new synchronized sequence could be rendered. Dynamic Time Warping with a 10 frames window was used and highly accurate synchronization was achieved, Figure 3 shows the precise synchronization using the proposed approach. The accuracy of the result is far more impressive when viewed as a video. The sequence is aligned according to the actions of the left-most actor.

Figure 3. Following the leader. The top row shows six frames before synchronization. Notice the difference in postures of each actor in a single view. The bottom row shows corresponding frames (to the top row) from the rendered sequence after synchronization.

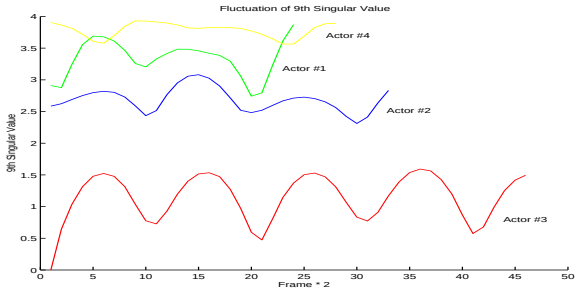

Figure 4. Finding the Odd Man Out. Actor three corresponds to the actor performing the 'Egyptian' gait.

	1-1	1-2	2-1	2-2	3-1	3-2
1-1	0.000	1.152	1.862	2.131	1.439	1.581
1-2	1.152	0.000	1.963	2.325	1.498	1.568
2-1	2.014	2.611	0.000	0.870	1.468	1.649
2-2	2.222	2.985	0.870	0.000	1.541	1.739
3-1	1.443	1.615	1.760	2.119	0.000	1.106
3-2	1.667	1.928	2.530	2.611	1.106	0.000

Table 1. Distance Matrix for Gait Analysis. The notation 1-1 refers to 'Actor 1, View 1' etc. Note that lower values correspond to the same actor's gaits in different views (1-1 matches best with 1-2, 2-1 with 2-2, 3-1 with 3-2).

4.3 Gait Analysis

Three walking actors were captured from two different view points using two cameras, and, on average, each video was more than 200 frames in length. Six feature points, hands, knees and feet, were tracked. An arbitrary short fragment (40 frames) was extracted from each video. The goal of experiment was determining if the extracted fragment could be found in the videos and computing the ninth singular value as the best similarity measure. Table 4.3 shows the distance matrix of each gait in each view. In the table the first and second columns correspond to the first actor in the first and second view respectively, and so on. As expected, the distance between the gait of an actor in first view and in the second view is always lower than the gait of other actors in any view.

5 Conclusion

The objective of this work has been to find an approach to matching human actions that is both fully descriptive in terms of motion and is both invariant to view and execution rate. To our knowledge, this is the first work that expressly addresses the variability of human proportions. We make innovative use of epipolar geometry to propose a similarity measure between two sets of actions. Two related constraints were proposed and explored. Experimental validation of the proposed approach show the versatility and importance of the results obtained in this work. In this work, we assume that the anatomical landmarks are identified and that they have been tracked for the duration of the analysis. Future work includes extending the proposed approach to handle more challenging situations like partial occlusions, erroneous tracks etc, and using statistical models for estimating membership to classes of actions.

References

[1] N. Badler, C. Philips, and B. Webber. *Simulating Humans*. Oxford University Press, 1993.

[2] R. Bridger. *Human Performance Engineering: A Guide For System Designers*. Prentice-Hall, 1982.

[3] R. Bridger. *Introduction to Ergonomics*. McGraw-Hill, 1995.

[4] R. Easterby, K. Kroemer, and D. Chaffin. *Anthropometry and Biomechanics - Theory and Application*. Plenum Press, New York, 1982.

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2003.

[6] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, 2000.

[8] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1993.

[9] J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[10] R. Rosales and S. Sclaroff. Trajectory guided tracking and recognition of actions. *PAMI Special Issue on Video Surveillance and Monitoring*, 1999.

[11] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.ASSp-26*, pages 43–49, 1978.

[12] I. M. Unknown. Reference removed.

[13] L. Zelnik-Manor and M. Irani. Event-based video analysis. *CVPR*, 2001.