Recovering Articulated Model Topology from Observed Motion

Leonid Taycher, John W. Fisher III, and Trevor Darrell
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, 02139
{lodrion, fisher, trevor}@ai.mit.edu

Abstract

Tracking human motion is an integral part to developing powerful human-computer interfaces. Several successful tracking algorithms were developed that model human body as an articulated tree. We propose a learning-based method for creating such articulated models from observations of multiple rigid motions. This paper is concerned with recovering topology of the articulated model, when the rigid motion of constituent segments is known.

Our approach is based on finding the maximum likelihood tree shaped factorization of the joint probability density function (PDF) of rigid segment motions. The topology of graphical model formed from this factorization corresponds to topology of the underlying articulated body. We demonstrate the performance of our algorithm both on synthetic and real motion capture data.

1. Introduction

Full-body tracking and analysis of biological motion have become active research topics in recent years. A common approach to this task is to model the body as a kinematic tree, and reformulate the problem as an articulated body tracking[8]. Most of the state-of-the-art systems rely on predefined kinematic models [21, 20, 18]. Some methods require manual initialization, while other use heuristics [15, 7], or predefined protocols [13] to adapt the model to observations.

We are interested in a principled way to recover articulated models from observations. The recovered models may then be used for further tracking and/or recognition. We would like to approach model estimation as a multistage problem. In the first stage the rigidly moving segments are tracked independently; at the second stage, the topology of the body (the connectivity between the segments) is recovered. After the topology is determined, the joint positions and the joint angle limits can be determined.

In this paper we concentrate on the second stage of this task, estimating the underlying topology of the observed articulated body, when the motion of the constituent rigid bodies is known. We approach this as a learning problem, in the spirit of [19]. If we assume that the body may be modeled as a kinematic tree, and motion of a particular rigid segment is known, then the motions of the rigid segments that are connected through that segment are independent of each other. That is, we can model a probability distribution of the full body-pose as a tree-structured graphical model, where each node corresponds to pose of a rigid segment. This observation allows us to formulate the problem of recovering topology of an articulated body as finding the tree-shaped graphical model that best (in Maximum Likelihood sense) describes the observations.

The rest of the paper is structured as follows: in Section 2 we describe relevant prior work, we then describe the probabilistic formulation in Section 3, and finally we present the algorithm used for computations (Section 4), our experiments and the conclusions.

2. Prior Work

While state-of-the-art tracking algorithms [21, 9, 5, 20, 18] do not address either model creation or model initialization, the necessity of automating these two steps has been long recognized.

The approach in [13] required a subject to follow a set of predefined movements, and recovered the descriptions of body parts and body topology from deformations of apparent contours. Various heuristics were used in [15, 7] to adapt an articulated model of known topology to 3D observations. Analysis of magnetic motion capture data was used by [16] to recover limb lengths and joint locations for known topology, it also suggested similar analysis for topology extraction. A learning based approach for decomposing a set of observed marker positions and velocities into sets corresponding to various body parts was described in [19]. Our work builds on the latter two approaches in estimating the topology of the articulated tree model underlying the observed motion.

Several methods have been used to recover multiple rigid motions from video, such as factorization [3, 22], RANSAC [10], and learning based methods [12]. In this work we as-

sume that the 3-D rigid motions has been recovered and are represented using 2-D Scaled Prismatic Model (SPM).

2.1 Scaled Prismatic Model

A 2-D Scaled Prismatic Model (SPM) can be obtained by orthographically "projecting" 3-D model to the image plane[17]. An SPM has four degrees of freedom: in-plane translation, rotation, and uniform scale. 3-D rigid motion of an object, may be simulated by SPM transformations, using in-plane translation for rigid translation, rotation for uniform scaling for plane-parallel and out-of plane rotations respectively.

SPM motion (or pose) may be expressed as a linear transformation in projective space ($\mathfrak{P}^2$) as

$$\mathbf{M} = \begin{pmatrix} a & -b & e \\ b & a & f \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

If motion is not a pure translation, we can parameterize it by coordinates of fixed point (c_x, c_y), rotation angle α, and scale s,

$$\mathbf{M} = \begin{pmatrix} s\cos\alpha & -s\sin\alpha & c_x - sc_x\cos\alpha + sc_y\sin\alpha \\ s\sin\alpha & s\cos\alpha & c_y - sc_x\sin\alpha - sc_y\cos\alpha \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

3. Probabilistic Formulation

As previously stated, we wish to infer the underlying topology of an articulated body from noisy observations of a set of rigid body motions. Towards that end we will adopt a statistical framework for fitting a joint probability density over the observations. Here we describe the principles underlying our approach in very general terms. As a practical matter, one must make choices regarding density models underlying the observations. We discuss one such choice although other choices are also suitable.

We denote the set of observed motions of N rigid bodies at time $t, 1 \leq t \leq F$ as a set $\left\{\mathbf{M_s^t} | 1 \leq s \leq N\right\}$. Graphical models provide a useful methodology for expressing the dependency structure of a set of random variables(cf. [11]). Variables are assigned to the vertices of a graph, that is $M_i = \left\{\mathbf{M_i^t} | 1 \leq t \leq F\right\}$) while edges between nodes indicate dependency. We shall denote an edge between two variables, M_i and M_j by

$$E_{ij} = \begin{cases} 1 & \text{there is an edge between } M_i \text{ and } M_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Furthermore, if the corresponding graphical model is a spanning tree, it can be expressed as a product of conditional densities (e.g. see [14])

$$P_M\left(M_1, \ldots, M_N\right) = \prod_{M_s} P_{M_s|\text{pa}(M_s)}\left(M_s|\text{pa}\left(M_s\right)\right) \tag{4}$$

where pa(M_s) is the parent of M_s. While multiple nodes may have the same parent, each individual node has only one parent node. Furthermore, in any decomposition one node (the root node) has no parent. Any node (variable) in the model can serve as the root node [11]. Consequently, a tree model puts constraints on E. Of the possible tree models (choices of E), we wish to choose the maximum likelihood tree which is equivalent to the minimum entropy tree [4]. The entropy of a tree model can be written

$$H(M) = \sum_s H(M_s) - \sum_{E_{ij}=1} I(M_i; M_j) \tag{5}$$

where $H(M_s)$ is the marginal entropy of each variable and $I(M_i; M_j)$ is the mutual information between nodes M_i and M_j and quantifies their statistical dependence. Consequently, the minimum entropy tree corresponds to the choice of E which minimizes the sum of the pairwise mutual informations [1]. The tree denoted by E can be found via the maximum spanning tree algorithm [2] using $I(M_i; M_j)$ for all i, j as the edge weights.

Our conjecture is that the if our data are sampled from a variety of motions the topology of the estimated density model is likely to be the same as the topology of the articulated body model. The follows from the intuition that when considering only pairwise relationships, the relative motions of physically connected bodies will be most strongly related.

3.1 Estimation of Mutual Information

As a necessary step to choosing the minimum entropy spanning tree we must estimate the pairwise mutual informations between rigid motions M_i and M_j for all i, j pairs. As stated, in order to do so we must make a choice regarding the parameterization of motion and a probability density over that parameterization. Since in this work we are concerned with extracting articulated model topology and not the 3-D descriptions of the model, we have elected to use Scaled Prismatic Model (Section 2.1),

We parameterize rigid motions, $\mathbf{M_i^t}$, by the vector of quantities $m_i^t = (c_x, c_y, s, \alpha)^T$ where (c_x, c_y) is the instantaneous center of rotation, s is a relative scaling, and α is the rotation. In general,

$$H(\mathbf{M_i}) \neq H(m_i) \tag{6}$$

but, since there is a one-to-one correspondence between the $\mathbf{M_i}$'s and m_i's [4]

$$I(\mathbf{M_i};\mathbf{M_j}) = I(m_i;m_j) \tag{7}$$

and consequently we can estimate the $I(\mathbf{M_i};\mathbf{M_j})$ by first computing m_i^t, m_j^t from $\mathbf{M_i^t}, \mathbf{M_j^t}$. From the normal distribution assumption, the pairwise mutual informations, $I(m_i;m_j)$, are a function of the estimated covariances matrices and can be computed thusly

$$\begin{aligned} I(m_i;m_j) &= H(m_i) + H(m_j) - H(m_i,m_j) \quad (8)\\ &= \frac{1}{2}\log\left((\pi e)^4 \left|\Sigma_{m_i}\right|\right) + \\ &\quad \frac{1}{2}\log\left((\pi e)^4 \left|\Sigma_{m_j}\right|\right) - \\ &\quad \frac{1}{2}\log\left((\pi e)^8 \left|\Sigma_{m_i,m_j}\right|\right) \quad (9) \end{aligned}$$

or equivalently

$$\begin{aligned} I(m_i;m_j) &= H(m_i) - H(m_j|m_i) \quad (10)\\ &= \frac{1}{2}\log\left((\pi e)^4 \left|\Sigma_{m_i}\right|\right) - \\ &\quad \frac{1}{2}\log\left((\pi e)^4 \left|\Sigma_{m_j|m_i}\right|\right) \quad (11) \end{aligned}$$

where $m_j|m_i$ indicates a relative motion described in the next section. In practice, the estimates of covariance matrices are not perfect and furthermore the Gaussian assumption is only an approximate model. Consequently, some estimates of mutual information may yield invalid results. These terms are set to zero in practice (no edge will be placed between these nodes). Note that the Gaussian model on the m_i's does not assume Gaussianity on the M_i's due to their nonlinear relationship. Furthermore we approximate $m_j|m_i$ with $m_{j|i}$ derived from equation 13.

4. Algorithm

The input to our algorithm is a set of SPM poses (Section 2.1) $\{\mathbf{P_s^t}|1 \le s \le S, 1 \le t \le T\}$, where S is the number of rigid segments tracked and F is the number of frames. In order to compute the mutual information between the motion of segments s_1 and s_2, we first compute motions of segment s_1 in frames $1 < t \le F$ relative to its position is frame $t_1 = 1$,

$$\mathbf{M_{s_1}^{t_1 t}} = \mathbf{P_{s_1}^t}(\mathbf{P_{s_1}^{t_1}})^{-1} \tag{12}$$

and transformation of s_1 relative to s_2 (with the relative pose $\mathbf{P_{s_1|s_2}} = (\mathbf{P_{s_2}})^{-1}\mathbf{P_{s_1}}$),

$$\mathbf{M_{s_1|s_2}^{t_1 t}} = ((\mathbf{P_{s_2}^t})^{-1}\mathbf{P_{s_1}}^t)((\mathbf{P_{s_2}^{t_1}})^{-1}\mathbf{P_{s_1}^{t_1}})^{-1} \tag{13}$$

The parameter vectors $m_{s_1}^{t_1 t} = (c_x, c_y, s, \alpha)^T$ and $m_{s_1|s_2}^{t_1 t}$ are then extracted from the transformation matrices $\mathbf{M_{s_1}}$ and $\mathbf{M_{s_1|s_2}}$ (cf. Section 2.1), and the mutual information is estimated as described in Section 3.1. In order to avoid numerical errors in estimating parameters (and propagating them to mutual information computation), we disregard any frames t for which either $\alpha_{s_1}^{t_1 t} < \pi/12$ or $\alpha_{s_1|s_2}^{t_1 t} < \pi/12$, since estimating coordinates (c_x, c_y) of the instantaneous center of rotation is numerically unstable for motions with small rotations.

5. Results

We have tested our algorithm both on synthetic and motion capture data. Two synthetic sequences were generated in the following way. The rigid segments were positioned by randomly perturbing parameters of the corresponding kinematic tree structure. A set of feature points was then selected for each segment. At each time step point positions were computed based on the corresponding segment pose, and perturbed with Gaussian noise with zero mean and standard deviation of 1 pixel. The inputs to the algorithm were the segment poses re-estimated from the feature point coordinates. In the motion capture-based experiment, the segment poses were estimated from the marker positions.

The results of the experiments are shown in the Figures 5.1, 5.2 and 5.3. The first experiment involved a simple kinematic chain with 3 segments in order to demonstrate the operation of the algorithm. The sample configurations of the articulated body are shown in the first row of the Figures 5.1. The poses of the middle and right segments relative to the left one are shown in the next two rows. As can be seen from comparing the second and third row, knowledge about pose of the left segment provides much more information about the middle segment than about the right one (the motion of middle segment relative to the left one is a pure rotation). The graph computed using method from Section 3.1 and the corresponding maximum spanning tree are in Figures 5.1(m, o).

The second experiment involved a humanoid torso-like synthetic model containing 5 rigid segments. It was processed in a way similar to the first experiment. The results are shown in Figure 5.2.

For the human motion experiment, we have used motion capture data of a dance sequence (Figure 5.3(a-d)). The rigid segment motion was extracted from the positions of the markers tracked across 220 frames (the marker correspondence to the body locations was known). The algorithm was able to correctly recover the articulated body topology (Compare Figures 5.3(f) and 5.3(a)), when provided only with the extracted segment poses. The dance is a highly structured activity, so not all degrees of freedom were explored in this sequence, and mutual information be-

tween some unconnected segments (e.g. thighs S_3 and S_7) was determined to be relatively large, although this did not impact the final result.

6. Conclusions and Future Work

We have presented a novel general technique for recovering the underlying articulated structure from information about rigid segment motion under very weak assumptions (that this structure may be represented by a tree with unknown topology). While the results presented in this paper were obtained using the Scaled Prismatic model and Gaussian probability densities our methodology does not rely on either modeling assumption. Alternative parameterizations will be the subject of future analysis. The further extensions of this work would also include automatic localization of the joints between the neighboring segments in the articulated tree and determination of the degrees of freedom for each joint. Together with improved rigid segment tracking this would bring us close to solving an important task of automatic creation and initialization of models for articulated tracking.

References

[1] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, May 1968.

[2] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivern. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[3] Joao Paolo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[5] Jonathan Deutscher, Andrew Blake, and Reidm Ian. Articulated body motion capture by annealed particle filtering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[6] David E. DiFranco, Tat-Jen Cham, and James M. Regh. Reconstruction of 3-d figure motion from 2-d correspondences. In *Computer Vision and Pattern Recognition*, 2001.

[7] Dariu M. Gavrila and Larry S. Davis. Tracking of humans in action: a 3-d model-based approach. In *ARPA Image Understanding Workshop*, Palm Springs, Feb 1996.

[8] David C. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[9] Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Advances in Neural Information Processing Systems*, 12, 2000.

[10] Yi-Ping Hung, Cheng-Yuan Tang, Sheng-Wen Shin, Zen Chen, and Wei-Song Lin. A 3d feature-based tracker for tracking multiple moving objects with a controlled binocular head. Technical report, Academia Sinica Institute of Information Science, 1995.

[11] Finn Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[12] N. Jojic and B.J. Frey. Learning flexible sprites in video layers. In *Computer Vision and Pattern Recognition*, pages I:199–206, 2001.

[13] Ioannis A. Kakadiaris and Dimirti Metaxas. 3d human body acquisition from multiple views. In *Proc. Fifth International Conference on Computer Vision*, pages 618–623, 1995.

[14] Marina Meila. *Learning Mixtures of Trees*. PhD thesis, MIT, 1998.

[15] Ivana Mikic, Mohan Triverdi, Edward Hunter, and Pamela Cosman. Articulated body posture estimation from multicamera voxel data. In *Computer Vision and Pattern Recognition*, 2001.

[16] J. O'Brien, R. E. Bodenheimer, G. Brostow, and J. K. Hodgins. Automatic joint parameter estimation from magnetic motion capture data. In *Graphics Interface'2000*, pages 53–60, 2000.

[17] James M. Regh and Daniel D. Morris. Singularities in articulated object tracking with 2-d and 3-d models. Technical report, Digital Equipment Corporation, 1997.

[18] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proc. European Conference on Computer Vision*, 2000.

[19] Yang Song, Luis Goncalves, Enrico Di Bernardo, and Pietro Perona. Monocular perception of biological motion - detection and labeling. In *Proc. International Conference on Computer Vision*, pages 805–812, 1999.

[20] B. Stenger, P. R. S. Mendonca, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. *Proc. British Machine Vision Conference*, 2001.

[21] Ying Wu, Jonh Y. Lin, and Thomas S. Huang. Capturing natural hand articulation. In *Proc. International Conference on Computer Vision*, 2001.

[22] Ying Wu, Zhengyou Zhang, Thomas S. Huang, and John Y. Lin. Multibody grouping via orthogonal subspace decomposition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

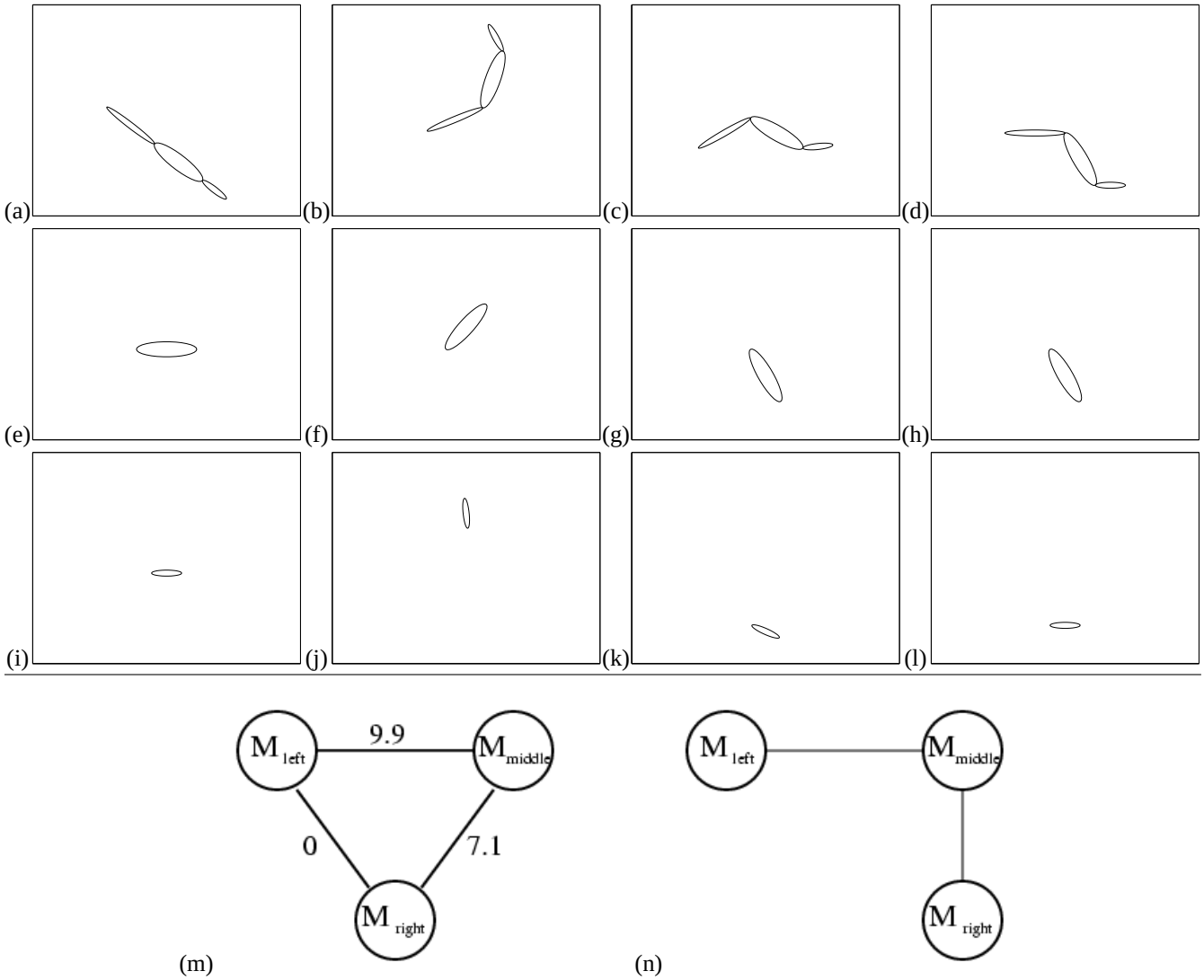

Figure 5.1: Simple kinematic chain topology recovery. The first row shows 4 sample frames from a 100 frame synthetic sequence. The next two rows show the poses of the middle and the right segments (respectively) relative to the left one. As can be seen, the pose of the left segment provides much more information about the pose of the middle segment (to which is it directly connected), than about the right one (the middle segment motion is a pure rotation). The mutual information graph is shown in (m), and the maximum spanning tree is in (n).

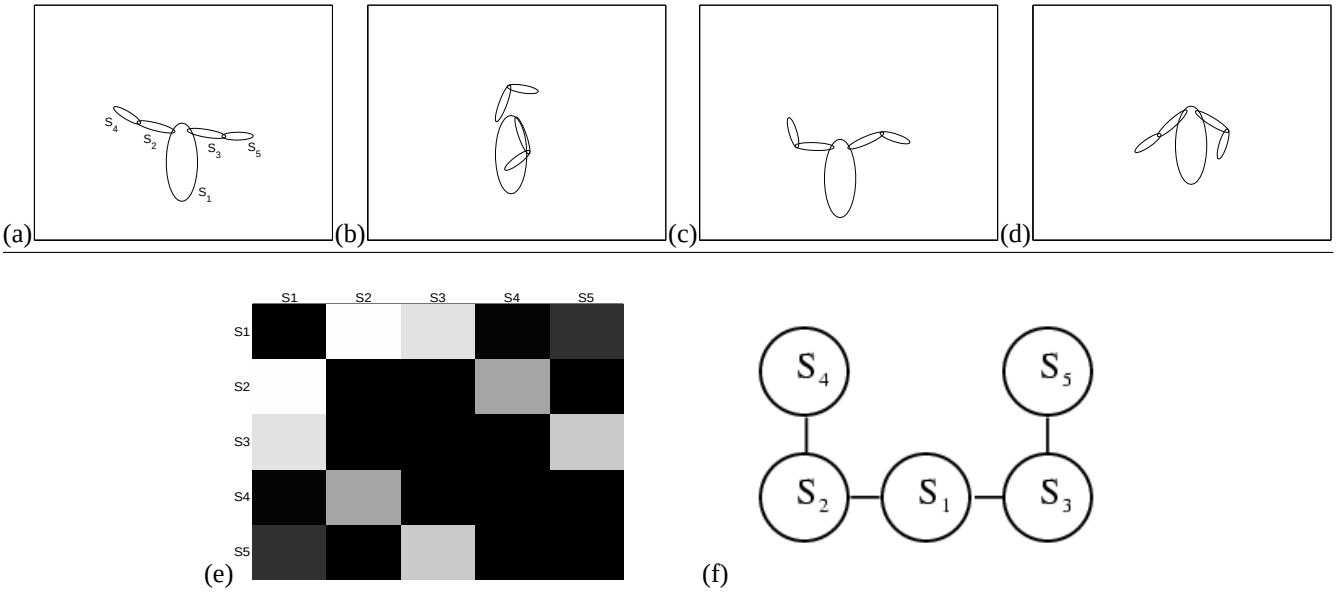

Figure 5.2: Humanoid torso synthetic test. The first sample frames from a randomly generated 150 frame sequence are shown in (a), (b), (c) and (d). The adjacency matrix of the mutual information graph is shown in (e), with intensities corresponding to edge weights. The vertices in the graph correspond to the rigid segments labeled in (a). (f) is the recovered articulated topology.

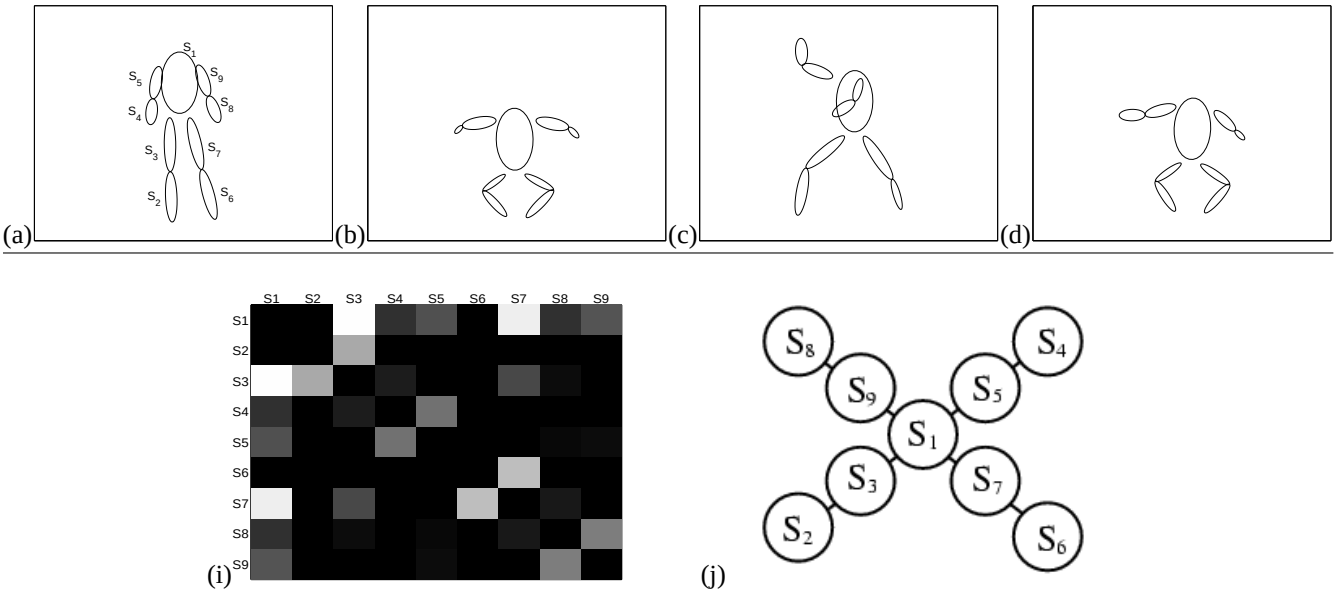

Figure 5.3: Motion Capture based test. (a), (b), (c) and (d) are the sample frames from a 220 frame sequence. The adjacency matrix of the mutual information graph is shown in (e), with intensities corresponding to edge weights.The vertices in the graph correspond to the rigid segments labeled in (a). (f) is the recovered articulated topology.