Towards Model-free Markerless Motion Capture

Chi-Wei Chu, Odest Chadwicke Jenkins, Maja J Matarić
Center for Robotics and Embedded Systems
Robotics Research Laboratories
Department of Computer Science
University of Southern California
Los Angeles, CA 90089-0781
chuc,cjenkins,mataric@usc.edu

Abstract

An approach for model-free markerless motion capture of humans is presented. This approach is centered on generating underlying nonlinear axes (or a skeleton curve) from a volume of a human subject. Human volumes are captured from multiple calibrated cameras. We describe the use of skeleton curves for determining the kinematic posture of a human captured volume. Our motion capture uses a skeleton curve, found in each frame of a volume sequence, to automatically produce kinematic motion. We apply several types of motion to our capture approach and use the results to actuate a dynamically simulated humanoid robot.

1 Introduction / Problem Statement

The ability to collect human motion data is an invaluable tool in controlling humanoid robots. This fact can be seen by the increasing usage of *motion capture* technologies for uses such as teleoperation, driving dynamically animated characters [18], and development of basis behavior for control [8]. Current motion capture techniques, however, suffer from several bothersome limitations, such as:

1. motion capture systems can be prohibitively expensive
2. the captured subject must be instrumented with (usually) cumbersome equipment
3. the capture instrumentation limits the area in which a subject can be captured
4. consistent labels for passive markers, subject to occlusion, are difficult to determine
5. the conversion of marker data into a usable form, such as kinematic motion, is very difficult

Solutions to many such issues are typically ”workarounds”, ”hacks” or other unpublished ”tricks of the trade”.

An emerging area of research involves uninstrumented capture of motion, or *markerless motion capture.* For markerless motion capture, human data are acquired through some passive sensing mechanism and then reconciled into kinematic motion. Several *model-based* markerless motion capture approaches have been proposed ([3], [11]).

In this paper, we introduce one solution for markerless motion capture of human subjects from multiple calibrated cameras. Our experience with this problem leads us to hypothesize that *model-free* approaches can be developed to address all or part of this problem. Towards this end, we have developed a model-free method, called *nonlinear spherical shells (NSS)* for extracting *skeleton point features* that are linked into a *skeleton curve* for a particular frame within a motion. A skeleton curve is an approximation of the ”underlying axis” of the human. The underlying axis is similar to the axis of a generalized cylinder or the wire spine of a poseable puppet. Using skeleton curves provided by NSS, we define a “volume-to-posture” procedure to automatically determine the kinematic posture of the human subject in each frame of a motion sequence.

Several advantages arise in using our approach for markerless motion capture. First, our method is fast and accurate enough to be tractably applied to all frames in a motion. Our method can be used alone or as an initialization step for model-based capture approaches. Second, our dependence on modeling human bodies decreases because we only use simple assumptions about the topology of human kinematics. Third, the posture of the human subject is automatically determined without complicated ”workarounds”.

We demonstrate the success of our system by capturing three different types of motions and using the captured motion to actuate a 20 DOF dynamically

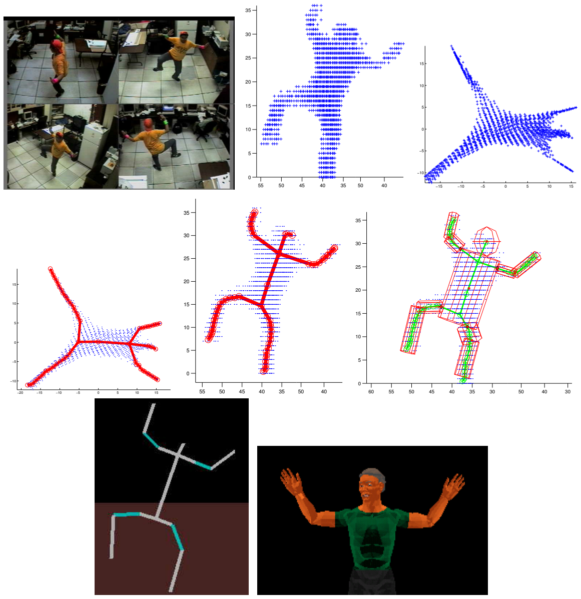

Figure 1: The outline of our appraoch. (1) A human viewed in multiple cameras is used to build (2) a Euclidean space point volume of the human. This volume is transformed into (3) an intrinsic space pose-invariant volume and its (4) principle curves are found. The principal curves are projected back into Euclidean space to provide (5) a skeleton curve. The skeleton curve is used to determine (6) the posture of the human. Using the postures of a volume sequence, (7) the kinematic motion of the human is found and (8) actuated on the Adonis humanoid simulation.

simulated humanoid upper body, Adonis [10].

2 Background / Hypothesis

We considered two types of previous approaches to markerless motion capture. The key difference between these approaches is in how features found from 2D image data of each camera are combined to produce information in a 3D world space. The first approach, including such techniques as [3], [5], [2], uses 2D image features from each camera to directly search for (or update to) an appropriate configuration for a 3D body model. The other markerless capture approach introduces an intermediate step of geometrically building a 3D volume feature of the capture subject from 2D image features, using a technique such a voxel carving ([15], [12]. A 3D body model is then fit or updated to fit the current 3D volume, as in [11].

None of these previous approaches are oriented towards model-free markerless motion capture. These capture approaches require either:

1. an appropriately constrained search in a vast configuration space of a known human body model or
2. a temporal dependence between frames to allow previously determined postures of a known human body model to guide the update of the model to its current configuration

We believe that the need for an a priori human body model can be diminished or eliminated by leveraging if the nonlinear structure of the 3D volume feature can be found. Tracking procedures, such as [11], perform model initialization and tracking on a voxel volume feature expressed in 3D Euclidean space. The body models used in this type of procedure are a set of rigid body parts with linear central axes (such as a cylinder) connected by revolute joints. Nonlinearity is introduced into the model by the degrees of freedom provided by the revolute joints. By eliminating the nonlinearity induced by the joints, the body model can be placed into a pose-invariant posture, such as a "Da Vinci" posture.

Fortunately for us, recent work on manifold learning techniques have produced methods capable of uncovering nonlinear structure from spatial data. These techniques include Isomap [17], Kernel PCA [14], and Locally Linear Embedding [13]. Isomap works by building geodesic distances between data point pairs on an underlying spatial manifold. These distances are used to perform a nonlinear PCA-like embedding to an *intrinsic space*, a subspace of the original data containing the underlying spatial manifold. Isomap, in particular, has been demonstrated to extract meaningful nonlinear representations for high dimensional data such as images of handwritten digits, natural hand movements, and a pose-varying human head.

In this paper, we hypothesize that *the application of Isomap can transform a set of 3D points comprising a human point volume feature into a pose-invariant intrinsic space posture*, assuming distinguishable body limbs. By having such a transformation, we have a correspondence between volume points in both Euclidean and intrinsic spaces. We further hypothesize that features such as principal curves [6] can be trivially built in intrinsic space and mapped back to the volume feature in Euclidean space to produce a skeleton curve. Both of these hypothesizes assume the absence of an a priori body model.

3 Volume Capture

We describe an existing volume capture technique using multiple calibrated cameras. While not the focus of our work, this implementation does provide an adequate means for collecting input volume data. This implementation is derived from the work of Penny et. al. [12] for real-time volume capture; however, several other approaches are readily available [15]. In our capture setup, we place multiple cameras around three sides of a hypothetical rectangular volume, such that each camera can view roughly all of the volume. This rectangular volume is a voxel grid that divides the space in which moving objects can be captured.

The intrinsic and extrinsic calibration parameters for the cameras are extracted using Camera Calibration Toolbox designed by [1]. The parameters from calibration allow us to precompute a look-up table for mapping a voxel to pixel locations in each camera. For each frame in the motion, silhouettes of foreground objects in the capture space are segmented within the image of each camera and used to carve the voxel grid. A background subtraction method proposed in [4] was used. It can then be determined if each voxel in the grid is part of a foreground object by counting and thresholding the number of camera images in which it is part of a silhouette. One set of volume data is collected for each frame (i.e., set of synchronized camera images) and stored for offline processing.

4 Nonlinear Spherical Shells

Nonlinear spherical shells (NSS) is our model-free approach for building a skeleton curve feature for a Euclidean space volume of points. The procedure for (NSS) works in three main steps:

1. Removal of pose-dependent nonlinearities from the volume by transforming the volume into an intrinsic space using Isomap

2. Dividing and clustering the pose-independent volume such that principal curves are found in intrinsic space

3. Project points defining the intrinsic space principal curve into the original Euclidean space to produce a skeleton curve for the point volume

Isomap is applied in the first step of the NSS procedure to remove pose nonlinearities from a set of points compromising the captured human in Euclidean space. We use the implementation provided by the authors of Isomap, which is available at http://isomap.stanford.edu/. This implementation of Isomap is applied directly to the volume data. Isomap requires the user to specify only the number of dimensions for the intrinsic space and how to construct local neighborhoods for each data point. Dimension reduction is not our aim. Thus, the intrinsic space is set to have 3 dimensions. Each point determines other points within its neighborhood using k-nearest neighbors or an epsilon sphere with a chosen radius.

The application of Isomap transforms the volume points into a pose-independent arrangement in the intrinsic space. The pose-independent arrangement is similar to a "da Vinci" pose in 3 dimensions (figure 2). Isomap can produce the da Vinci point arrangement for any point volume with distinguishable limbs.

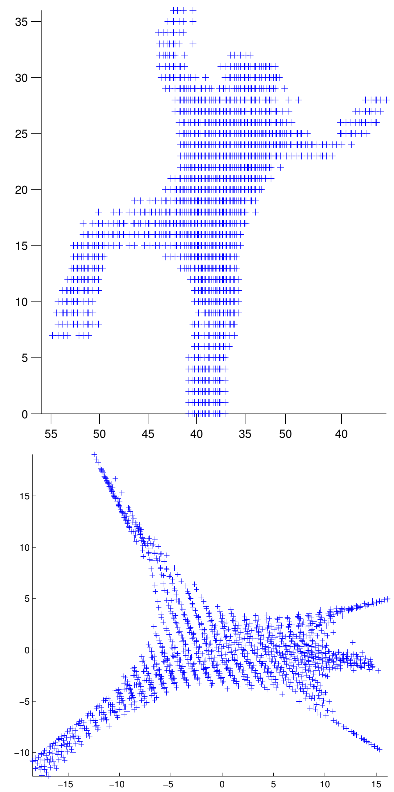

Figure 2: *A captured human volume in Euclidean space (left) and its pose-invariant intrinsic space representation (right).*

The next step in the NSS procedure is processing intrinsic space volume for principal curves. The definition of principal curves can be found in [6] or [9] as "self-consistent" smooth curves which pass through the "middle" of a d-dimensional data cloud, or nonlinear principal components. While smoothness is not our primary concern, we are interested in placing a curve through the "middle" of our Euclidean space volume. Depending on the posture of the human, this task can be difficult in Euclidean space. However, the pose-invariant volume provided Isomap makes the determination of principal curves simple, due to properties of the intrinsic space volume. Isomap provides an intrinsic space volume that is mean-centered at the origin and has limb points that extend away from the origin.

Points on the principle curves be found by the following subprocedure (figure 3):

1. partitioning the intrinsic space volume points into *concentric spherical shells*
2. clustering the points in each partition
3. averaging the points of each cluster to produce a principal curve point
4. linking principal curve points with overlapping clusters in adjacent spherical shells

Clustering used for each partition was developed from the one-dimensional "sweep-and-prune" techniques for bounding-box clustering without specifying the expected number of clusters.

The result from the principal curves procedure is a set of points defining the principal curves linked in a hierarchical tree structure. The root of the principal curves is the mean of the volume.

The final step in the NSS procedure projects the intrinsic space principal curve points into a skeleton curve in the original Euclidean space. We use Shepard's interpolation [16] to map principal curve points onto the Euclidean space volume to produce skeleton curve points. The skeleton curve is formed by applying the same tree-structured linkages as the intrinsic space principal curves to the skeleton curve points.

4.1 Skeleton Curve Refinement

The skeleton curve found by the NSS procedure will be indicative of the underlying spatial structure of the Euclidean space volume, but may contain a few undesirable artifacts. We handle these artifacts using a skeleton curve refinement procedure. The refinement procedure first eliminates *noise branches* in the skeleton curve that typically occur in areas of articulation, such as the hands, shoulders, feet, etc. Noise branches are detected as branches with depth under some threshold. Noise branches eliminated through merging its skeleton curve points with a non-noise branch.

The refinement procedure then eliminates noise for the root of the skeleton curve. Shell partitions around the mean of the body volume will be encompassed by the volume, i.e. contain a single cluster spread across the shell. The skeleton curve points for such partitions will be roughly located near the volume mean. These skeleton curve points are merged to yield a new root to the skeleton curve. The result is a skeleton curve having a root and two immediate descendants, for the upper and lower sections of the body.

The upper and lower body are separated into limbs by finding branching nodes (i.e., the chest and the pelvis) for the skeleton curve. Branching nodes are formed by merging descendants of the skeleton curve root and ascendents of the roots of the individual limbs. The root of an individual limb is found by looking for subtrees of the skeleton curve that contain a leaf node and exhibit no branching.

5 Converting Volume Sequences Into Kinematic Motion

In this section, we describe the application of NSS within the context of a "volume-to-posture" technique for markerless motion capture. The volume-to-posture procedure automatically determines a joint angle configuration defining the posture of Euclidean space point volume. The volume-to-posture procedure is applied independently to each volume in a sequence, producing a specific kinematic model and posture for each volume. A second pass across the kinematic models for each volume is used to produce a single normalized kinematic model for the volume sequence.

The previously described NSS procedure is capable of producing skeleton curve features in a model-free fashion. To allow for the determination of posture, we introduce some model-based assumptions about the kinematic topology of the human. We believe the kinematic assumptions we introduce can be derived automatically. For future work, we will extend NSS to automatically derive kinematic models from volumes with better mechanisms for determining branching nodes and limb articulation nodes. For our current work, we incorporate simple assumptions about the topology of human kinematics to guide the inference of posture. We assume that a human kinematically consists of a root torso and five descendent limbs (one head, two two-link arms, and two two-link legs).

We use these assumptions, with the skeleton curve provided by NSS, to fit cylinders to each body in the kinematic hierarchy, except for a sphere used for the head. Before fitting body parts, the skeleton curve limbs are identified as the head (the shortest branch), an arm (a limb connected to the chest branching node), or a leg. Each limb and the torso has an associated set of volume points defined by the clustering results from NSS. The head is the first body part to fit using a sphere. The skeleton curve nodes of the head branch are merged into a single point defining the center of the head sphere. The head sphere is grown to encompass all voxels associated with the head limb.

Next, a cylinder is grown to fit the root torso. The vector between the chest and pelvis nodes is the axis of the cylinder. The height and radius are initially set to bound the voxels of the root torso. The top and bottom of the torso cylinder are then extended to meet the head sphere and the pelvis branching node. Voxels that fall within the torso cylinder are reassociated to the torso regardless of their previous association from clustering within NSS.

The remaining limbs are two-link articulations and require a slightly more sophisticated cylinder fitting routine. For these limbs, it is necessary to determine the point at which the limb articulates (i.e., the location of the elbow/knee joint). Because these limbs have only two links, we can look for the node that indicates the "bend" in the skeleton curve . The skeleton point node defining the bend has the maximum angle between vectors from limb root to limb leaf and from limb root to the node. Volume points are classified into the upper and lower parts of the limb based on the limb articulation point. Cylinders are then fit to the points of each limb part. If there is no significant bend in the limb, no separation is applied to the points, leaving joint angles to be determined in the second pass.

The cylinders fit to each body part of the human help define the joint angle posture of the human. A world space coordinate system is found for each cylinder using the cylinder axis as the Z axis. The root torso defines its X axis as the cross product between its Z axis and the vector between the root points of each arm. The X axis for all other body parts is the cross product between its parent's X axis and its own Z axis. A special case is the head sphere, with the head Z axis defined by the vector difference between the head sphere center and chest branching node. The world space coordinate system for each body is converted to a local coordinate system by determining its 3D rotational transformation from its parent. This 3D rotation provides the joint angle configuration for the current posture of the human.

Finally, a second pass across the kinematic models and postures bring temporal consistency to the motion sequence. The second pass finds correspondences between body parts in subsequent frames for a consistent, but not semantic, labeling (i.e. "arm1-upper" instead of "upper right arm"). Parameters for the cylinders and the head sphere are averaged to provide a normalized kinematic model for the motion sequence. This kinematic model is used to refit and find joint angles for limbs without significant articulation.

6 Results and Observations

In this section, we describe our implementation of our markerless motion capture approach and results from its application to three different motion sequences: waving, walking, and jumping jacks. We have implemented our approach for markerless motion capture using Microsoft Visual C++ for volume capture and Matlab for our volume-to-posture and nonlinear spherical shells procedures. The execution of the entire capture procedure was performed on a 350 MHz Pentium with 128 MB of memory.

For each motion sequence performed by a human subject, a volume sequence was captured and stored for offline processing by the volume-to-posture procedure. Using the Intel Image Processing Library, we were able to capture volumes within a $80 \times 80 \times 50$ grid of cubic $50mm^3$ voxels at 10 Hz. Each volume sequence consisted of roughly 50 frames. Due to our frugal choices for camera and framegrabber options, our ability to capture human volumes was significantly restricted. Our image technology allowed for 320×240 image data from each camera, which produced several artifacts such as shadow voxels and ghosting voxels. This limitation restricted our capture motions to exaggerated, but usable motion, where the limbs were very distinct from each other. Improving our proof-of-concept volume capture system with more and better cameras, lighting, and computer vision techniques will vastly improve our markerless motion capture system, without having to adjust the volume-to-posture procedure.

Using the captured volume sequences, our volume-to-posture mechanism was able to accurately determine appropriate postures for each volume without fail. We used the same user parameters for each motion, consisting of an Isomap epsilon-ball neighborhood of radius $(50mm^3)^{1/2}$ and 25 for the number of concentric sphere partitions. In addition to accurate postures, the derived kinematic model parameters for each sequence appropriately matched the kinematics of the human capture subject. However, a significant amount of noise occurred between subsequent frames in the produced motion sequence. Noise is typical for many instrumented motion capture systems and should be expected when independently processing frames for temporally dependent motion. We were able to clean up this noise to produce aesthetically viable motion using standard low pass filtering.

Motions were output to files in the Biovision BVH motion capture format. Figure 6 shows the kinematic posture output for each motion. The BVH files were then successfully read and actuated by a 20 DOF dynamically simulated humanoid upper body, Adonis

[10]. From our results for these three motions, we believe that motion data for behaviors currently of interest for humanoid robots, such as kicking a soccer ball or swinging a tennis racket [7], can be captured by our current system. More images and movies of our results are available at http://robotics.usc.edu/ cjenkins/markerless/.

In observing the performance of our markerless capture system, several benefits of our approach became evident. First, the relative speed of our volume-to-posture procedure made the processing of each frame of a motion tractable. Depending on the number of volume points, the elapsed time for producing a posture from a volume by our Matlab implementation ranged between 60 and 90 seconds, with approximately 90 percent of this time spent for Isomap processing. Further improvements can be made to our implementation to speed up the procedure and process volumes with increasingly finer resolution. Second, our implementation required no explicit model of human kinematics, no initialization procedure, and no optimization of parameters with respect to a volume. Our model-free NSS procedure produced a representative skeleton curve description of a human posture based on the geometry of the volume. Kinematic assumptions about humans were used with the skeleton curve to guide the determination of posture, but a kinematic model with explicit degrees of freedom was not specified. Lastly, the skeleton curve may be a useful representation of posture in and of itself. Rigid-body motion is often represented through typically model-specific kinematics. Instead, the skeleton curve may allow for an expression of motion that can be shared between kinematic models, for purposes such as imitation.

7 Issues for Future Work

Using our current work as a platform, we aim to improve our ability to collect motion data of humans in various scenarios. Motion data is critically important for other related projects, such as the derivation of basis behaviors from motion data [8]. Areas for further improvements to our capture approach include:

1. automatically deriving kinematic models and postures for capture subjects with arbitrary kinematic topologies, both tree-structured and cyclical
2. exploring connections between model-free methods for robust model creation and initialization and model-based methods for robust temporal tracking
3. extensions to Isomap for volumes of greater resolutions and faster processing of data
4. using better computer vision techniques for volume capture to extend the types motion that can be converted into kinematic motion.

8 Conclusion

We have presented our initial efforts for a model-free approach to markerless motion capture. In revisiting our hypotheses, we have shown the application of Isomap to volume data provides both the removal of pose-dependent nonlinearities and extractable skeleton curve features for a captured human volume. We propose an approach, nonlinear spherical shells, for extracting skeleton curve features from a human volume. This feature extraction is placed within the context of a larger approach for converting volume sequences into kinematic motion. Our approach was successfully applied to different types of motion and used to actuate a humanoid robot. Building on our capture approach, we believe that control of humanoid systems can greatly benefit from the collection of human motion data for a variety of behaviors in increasingly for less structured, less cumbersome, and less expensive scenarios.

References

[1] Jean-Yves Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.

[2] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conference on Computer Vision and Pattern Recoginition*, pages 568–574, 1997.

[3] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133, 2000.

[4] Alexandre R.J. Francois and Gérard G. Medioni. Adaptive color background modeling for real-time segmentation of video streams. In *Proceedings of the International on Imaging Science, Systems, and Technology*, pages 227–232, Las Vegas, Nevada, June 1999.

[5] D.M. Gavrila and L.S. Davis. 3d model-based tracking of humans in action: A multi-view approach. In *CVPR96*, pages 73–80, 1996.

[6] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[7] A. J. Ijspeert, J. Nakanishi, and S. Schaal. Trajectory formation for imitation with nonlinear dynamical systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2001)*, pages 752–757, 2001.

[8] O. C. Jenkins and M. J Matarić. Deriving action and behavior primitives from human motion data (sample reference). In *Intl. Conf. on Robotics and Automation (to appear)*, 2002.

[9] Balazs Kegl, Adam Krzyzak, Tamas Linder, and Kenneth Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):281–297, 2000.

[10] Maja J Matarić, Victor B. Zordan, and Z. Mason. Movement control methods for complex, dynamically simulated agents: Adonis dances the macarena. In *Autonomous Agents*, pages 317–324, 1998.

[11] Ivana Mikić, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Articulated body posture estimation from multi-camera voxel data. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001.

[12] Simon G Penny, Jeffrey Smith, and Andre Bernhardt. Traces: Wireless full body tracking in the cave. In *Ninth International Conference on Artificial Reality and Telexistence (ICAT'99)*, December 1999.

[13] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[14] Bernhard Scholkopf, Alex J. Smola, and Klaus-Robert Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[15] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 1067–1073, 1997.

[16] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the ACM national conference*, pages 517–524. ACM Press, 1968.

[17] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[18] Victor B. Zordan and Jessica K. Hodgins. Motion capture-driven simulations that hit and react. In *Proceedings of the ACM SIGGRAPH symposium on Computer animation*, pages 89–96. ACM Press, 2002.

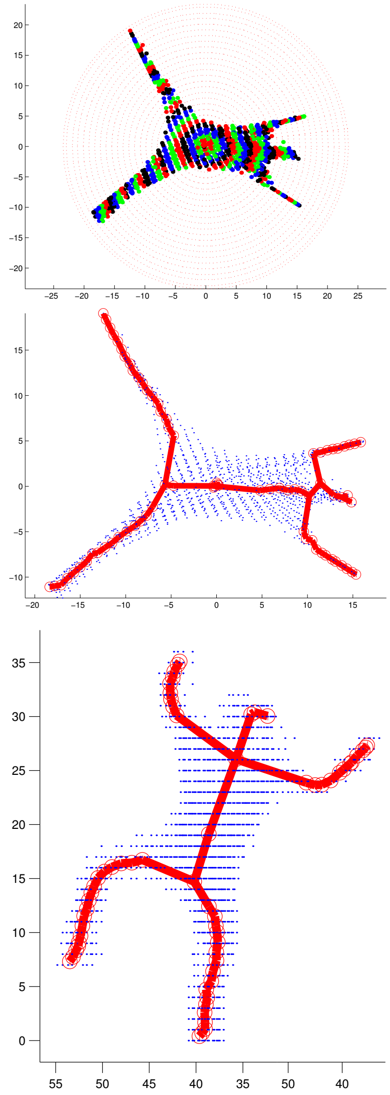

Figure 3: *Partitioning of the pose-invariant volume (top), its tree-structured principal curves (middle), and project back into Euclidean space (bottom).*

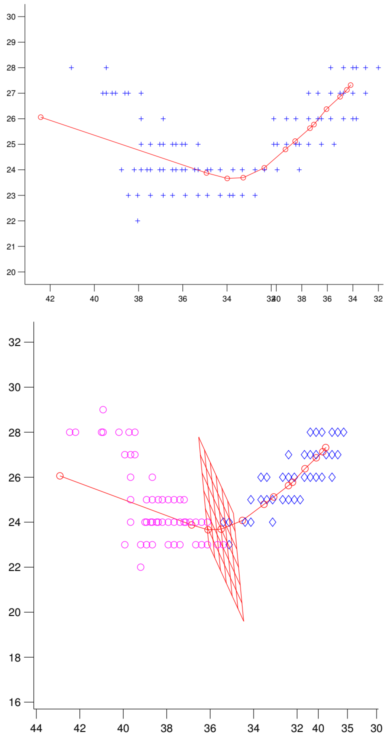

Figure 4: *Segmentation of an arm points into upper and lower parts.*

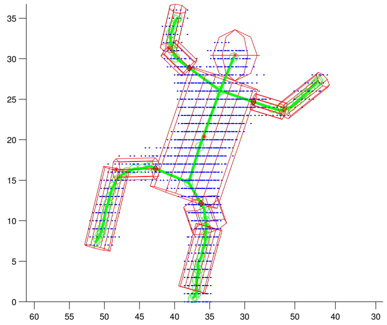

Figure 5: *The result from volume-to-posture procedure placing cylinders for each body part.*

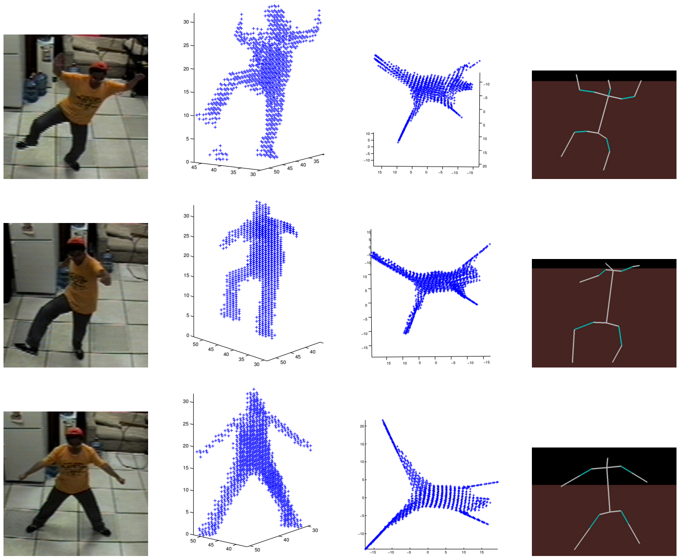

Figure 6: *Results from producing kinematic motion for waving, walking, jumping jacks (rows). The results are shown as a snapshot of the performing human, the capture point volume of the human, the pose-invariant volume and the derived kinematic posture (columns).*