Inferring 3D Structure with a Statistical Image-Based Shape Model

Kristen Grauman, Gregory Shakhnarovich, Trevor Darrell
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{*kgrauman, gregory, trevor*}*@ai.mit.edu*

Abstract

We present an image-based approach to infer 3D structure parameters using a probabilistic "shape+structure" model. The 3D shape of an object class is represented by sets of contours from silhouette views simultaneously observed from multiple calibrated cameras, while structural features of interest on the object are denoted by a number of 3D locations. A prior density over the multi-view shape and corresponding structure is constructed with a mixture of probabilistic principal components analyzers. Given a novel set of contours, we infer the unknown structure parameters from the new shape's Bayesian reconstruction. Model matching and parameter inference are done entirely in the image domain and require no explicit 3D construction. Our shape model enables accurate estimation of structure despite segmentation errors or missing views in the input silhouettes, and it works even with only a single input view. Using a training set of thousands of pedestrian images generated from a synthetic model, we can accurately infer the 3D locations of 19 joints on the body based on observed silhouette contours from real images.

1. Introduction

Estimating model shape or structure parameters from one or more input views is an important computer vision problem. Classic techniques attempt to detect and align 3D model instances within the image views, but high-dimensional models or models without well-defined features may make this type of search computationally prohibitive. Rather than fit explicit 3D models to input images, we explore reconstruction and parameter inference using image-based shape models which can be matched directly to observed features. We learn an implicit, image-based representation of a known 3D shape, match it to input images using a statistical model, and infer 3D parameters from the matched model.

Implicit representations of 3D shape can be formed using models of observed feature locations in multiple views. With sufficient training data of objects of a known class, a statistical multi-view appearance model can represent the most likely shapes in that class, as shown in [8]. Such a model can be used to reduce noise in observed images, or to fill in missing data.

In this paper we present an image-based method to infer 3D structure parameters using this sort of multi-view shape model. A probabilistic "shape+structure" model is formed using a probability density of multi-view silhouette contours augmented with 3D structure parameters (the 3D locations of key points on an object). We combine this with a model of the observation uncertainty of the silhouettes seen in each camera to compute a Bayesian estimate of an object's shape and structure. This estimate consists of the reconstructed multi-view contours (shape), and the inferred 3D parameters (structure). To our knowledge, this is the first work to formulate a multi-view statistical image-based shape model for the inference of 3D structure.

We also show how the image-based model can be learned from synthetic data, when available. Using a computer graphics model of articulated human bodies, we render a database of views augmented with the known 3D feature locations (and optionally joint angles, etc.) From this we learn a joint shape and structure model prior, which can be used to find the instance of the model class that is closest to a new input image. One advantage of a synthetic training set is that labeled real data is not required; the synthetic model includes 3D structure parameter labels for each example.

The strength of our approach lies in our use of a probabilistic multi-view shape model which restricts the object shape and its possible structural configurations to those that are most probable given the object class and the current observation. Even when given poorly segmented binary images of the object, the statistical model can infer appropriate structure parameters. Moreover, all computation is done within the image domain, and no model matching or search in 3D space is required.

In our experiments, we demonstrate how our shape+structure model enables accurate estimation of structure parameters despite large segmentation errors or even missing views in the input silhouettes. Since parameter inference with our model succeeds even with missing views, it is possible to match the model with fewer views than it has been trained on. In fact, in some cases we are able to get good parameter estimates with only

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

one input view. We also show how configurations that are typically ambiguous in single views are handled well by our multi-view model.

Our method has applications in many areas, including the fast approximation of 3D models for virtual reality applications, gesture recognition, pose estimation, and image feature correspondence matching across images.

2. Previous Work

In this paper we consider image-based statistical shape models that can be directly matched to observed image features. Models which capture the 2D distribution of feature point locations have been used to describe a wide range of flexible shapes, and they can be directly matched to input images [4]. The authors of [1] developed a single-view model of pedestrian contours, and showed how a linear subspace model formed from *principal components analysis* (PCA) could represent and track a wide range of motion [2]. A model appropriate for feature point locations sampled from a contour is also given in [2]. This single-view approach can be extended to 3D by considering multiple simultaneous views of features [8]. Shape models in several views can be separately estimated to match object appearance [5]; this approach was able to learn a mapping between the low-dimensional shape parameters in each view.

With multi-view contours from cameras at known locations, a visual hull can be recovered to model the shape of the observed object [11]. By forming a statistical model of these multi-view contours, an implicit shape representation that can be used for efficient reconstruction of visual hulls is created [8].

Our model is based on a mixture model whose components are estimated using PCA. The use of linear subspaces estimated by PCA to represent an object class, and more generally an appearance model, has been developed by several authors [3, 10, 16]. A probabilistic interpretation of PCA-based manifolds has been introduced by [9, 17] as well as in [13], where it was applied directly to face images. As described below, we rely on the mixture of *probabilistic principal components analyzers* (PPCA) formulation of [15] to model prior densities.

The idea of augmenting a PCA-based appearance model with structure parameters and using projection-based reconstruction to fill in the missing values of those parameters in new images was first proposed in [6]. A method that used a mixture of PCA approach to learn a model of single contour shape augmented with 3D structure parameters was presented in [14]; they were able to estimate 3D hand and arm locations just from a single silhouette. This system was also able to model contours observed in two simultaneous views, but separate models were formed for each so no implicit model of 3D shape was formed.

3. Bayesian Multi-View Shape Reconstruction

While regularization or Bayesian *maximum a posteriori* (MAP) estimation of single-view contours has received considerable attention as described above, less attention has been given to multi-view data from several cameras simultaneously observing an object. With multi-view data, a probabilistic model and MAP estimate can be computed on implicit 3D structures. We apply a PCA-based probability model to form Bayesian estimates of multi-view contours, and show how such a representation can be augmented and used for inferring structure parameters. Our work builds on the shape model introduced in [8], where a multi-view contour density model is derived for the purpose of 3D visual hull reconstruction.

Silhouette shapes are represented as sampled points on closed contours, with the shape vectors for each view concatenated to form a single vector. That is, with a set of n contour points $\mathbf{c}_k$ in each of the K views,

$$\mathbf{c}_k = (\mathbf{x}_1^k, \mathbf{x}_2^k, ..., \mathbf{x}_n^k), \quad 1 \leq k \leq K, \tag{1}$$

the $2Kn$-dimensional multi-view observation $\mathbf{o}$ is defined as

$$\mathbf{o} = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K)^T. \tag{2}$$

As described in [8], if the vector of observed contour points of a 3D object resides on a linear manifold, then the affine projections of that shape also belong to a linear manifold, at least for the case of affine cameras. Therefore, the shape vectors may be expressed as a linear combination of the 3D bases.

A technique suitable only for highly constrained shape spaces is to approximate the space with a single linear manifold. For more deformable structures, it is difficult to represent the shape space in this way. For example, with the pedestrian data we will use in the experiments reported below, inputs are expected to vary in two key (nonlinear) ways: the absolute direction in which the pedestrian is walking across the system workspace, and the phase of his walk cycle in that frame.

Thus, following [3, 15], we construct a density model using a mixture of PPCA models that locally models clusters of data in the input space with probabilistic linear manifolds. A single PPCA model is a probability distribution over the observation space for a given latent variable, which for this shape model is the true underlying contours in the multi-view image. Parameters for the M Gaussian mixture model components are determined for the set of observed data vectors $\mathbf{o}_n$, $1 \leq n \leq N$, using an EM algorithm to maximize a single log-likelihood function

$$L = \sum_{n=1}^{N} \log \sum_{i=1}^{M} \pi_i p(\mathbf{o}_n | i), \tag{3}$$

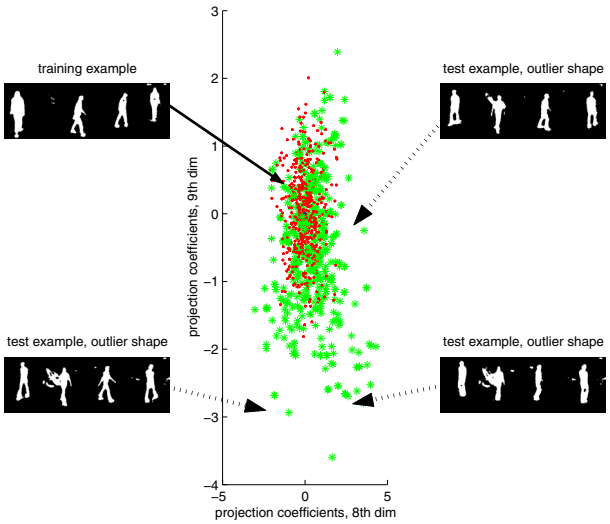


Figure 1: Illustration of prior and likelihood densities. Plot shows two projection coefficients in the shape subspace. The distribution of well-segmented silhouettes (dots) represents the prior shape density. The stars are novel observations. Due to large segmentation errors, some are unlikely samples according to the prior. MAP estimation reconstructs such contours as shapes closer to the prior.

where $p(\mathbf{o}_n|i)$ is a single PPCA model, and π_i is the i^{th} component's mixing proportion. A separate mean vector, principal axes, and noise variance is associated with each of the M components. As this likelihood is maximized, both the appropriate partitioning of the data and the respective principal axes are determined. The mixture of probabilistic linear subspaces constitutes the prior density of the object shape.

We assume there is a normal distribution of camera noise or jitter that affects the observed contour point locations in the input images, and we model this as a multivariate Gaussian.

A MAP estimate of the silhouettes is formed based on the PPCA prior shape model and the Gaussian observation likelihood. The estimate is then backprojected into the multi-view image domain to generate the recovered silhouettes. By characterizing which projections onto the subspace are more likely, the range of possible reconstructions is effectively moderated to be more like those expressed in the training set (see Figure 1). See [8] for details on this multi-view shape reconstruction process.

4. Inferring 3D Structure

The main contribution of this paper is the extension of the shape model described above to incorporate additional structural features. A model built to represent the shape of a

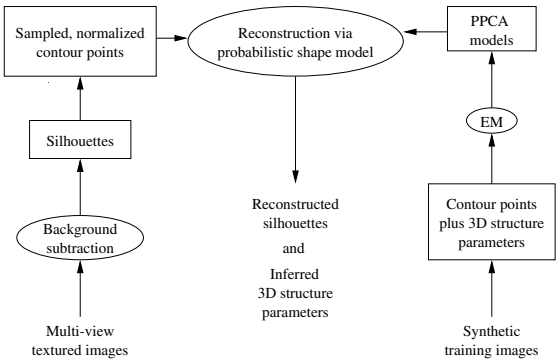


Figure 2: Diagram of data flow in our system.

certain class of objects using multiple contours can be augmented to include information about the object's orientation in the image, as well as the 3D locations of key points on the object. The mixture model now represents a density over the observation space for the true underlying contours together with their associated 3D structure parameters. Novel examples are matched to the contour-based shape model, using the same multi-view reconstruction method described in Section 3, in order to infer their unknown or missing parameters. See Figure 2 for a diagram of data flow.

The shape model is trained on a set of vectors composed of points from multiple contours from simultaneous views, plus a number of 3D structure parameters, $\mathbf{s}_j = (s_j^1, s_j^2, s_j^3)$. The $2Kn$+$3z$-dimensional observation vector $\mathbf{o}$ is then defined as

$$\mathbf{o} = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K, \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_z)^T, \quad (4)$$

where there are z 3D points. When presented with a new multi-view contour, we essentially treat the unknown 3D structure parameters as missing variables, and find the MAP estimate of the shape and structure parameters based on only the observable contour data. The training set for this inference task may be comprised of real or synthetic data, provided it has been labelled with the appropriate 3D structure parameters.

One strength of the proposed approach for the estimation of 3D feature locations is that the silhouettes in the novel inputs need not be cleanly segmented. Since the contours and unknown parameters are reconstructed concurrently, the parameters are essentially inferred from a restricted set of feasible shape reconstructions; they need not be determined by an explicit match to the raw observed silhouettes. Therefore, the probabilistic shape model does not require a perfect segmentation module. A fast, simple background subtraction scheme is sufficient.

As should be expected, our parameter inference method also benefits from the use of multi-view imagery. Multiple views will in many cases overcome the ambiguities that are geometrically inherent in single-view methods.

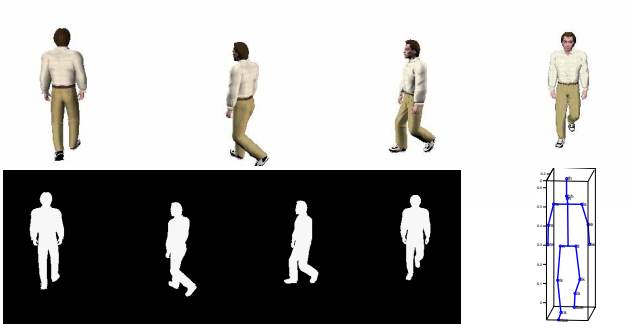


Figure 3: An example of synthetically generated training data. Textured images (top) show rendering of a human model from four viewpoints; silhouettes and stick figure (below) show multi-view contours and structure parameters, respectively.

5. Learning a Multi-View Shape Model for Pedestrians

A possible weakness of any shape model learned from examples is that the ability to accurately represent the space of realizable shapes will generally depend heavily on the amount of available training data. Moreover, we note that the training set on which the probabilistic shape+structure model is learned must be "clean"; otherwise the model could fit the bias of a particular segmentation algorithm. It must also be labeled with the true values for the 3D features. Collecting a large data set with these properties would be costly in resources and effort given the current state of the art in motion capture and segmentation, and at the end the "ground truth" could still be imprecise. We chose therefore to use realistic synthetic data for training a multi-view pedestrian shape model. We obtained a large training set by using POSER [7] – a commercially available animation software package – which allows us to manipulate realistic humanoid models, position them in the simulated scene, and render textured images or silhouettes from a desired point of view. Our goal is to train the model using this synthetic data, but then use the model for reconstruction and inference tasks with real images.

We generated 20,000 synthetic instances of multi-view input for our system. For each instance, a humanoid model was created with randomly adjusted anatomical shape parameters, and put into a walk-simulating pose, at a random phase of the walking cycle. The orientation of the model was drawn at random as well in order to simulate different walk directions of human subjects in the scene. Then for each camera in the real setup we rendered a snapshot of the model's silhouette from a point in the virtual scene approximately corresponding to that camera. In addition to the set of silhouettes, we record the 3D locations of 19 landmarks of the model's skeleton, corresponding to selected anatomical joints (see Figure 3).

For this model, each silhouette is represented as sampled points along the closed contour. All contour points are normalized to a translation and scale invariant input coordinate system, and each vector of normalized points is resampled to a common vector length using nearest neighbor interpolation. The complete representation is then the vector of concatenated multi-view contour points plus a fixed number of 3D body part locations (see Equation 4).

6. Experiments

We have applied our method to data sets of multi-view images of people walking. The goal is to infer the 3D positions of joints on the body given silhouette views from different viewpoints. In this section we first describe the experimental setup, the test data, and our error measures; we then summarize results from a variety of experiments and give some typical example outputs.

For the following experiments, we used an imaging model consisting of four monocular views per frame from cameras located at approximately the same height at known locations about 45 degrees apart. The working space of the system is defined as the intersection of their fields of view (approximately 3x3 meters). Images of subjects walking through the space at various directions are captured, and the silhouette foreground is extracted from each viewpoint.

The simple background subtraction algorithm used in our experiments is based on the notion of a static background that is occluded by a moving object; the implementation follows that in [12]. As a preprocessing step, a statistical model of the appearance of the background is built by collecting 400 images and calculating the mean and standard deviation of the graylevel for each pixel. The first stage marks as a candidate every pixel with a graylevel value which is more than three standard deviations away from the mean of that pixel (as learned from the background images). Next, normalized correlation with other candidate pixels in a small neighborhood around a candidate pixel is evaluated; this aims at removing some of the shadows, which cause large, but highly correlated changes with respect to the background. Finally, after applying a median filter in the neighborhood of a candidate pixel, the morphological "open" and "close" operations are applied to the candidate set. These operations, performed with small neighborhood size, are intended to remove small disconnected components and smooth the contours.

In these experiments, each view is a 320 by 240 image, and 200 points are sampled uniformly from each contour. In the input observation vector for each test example, the 3D pose parameters are set to zero. The number of mixture components M used is five.

Since we do not have ground truth pose parameters for

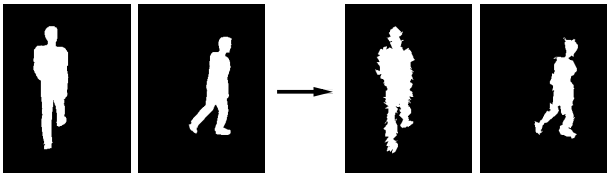


Figure 4: Two left images show clean synthetic silhouettes. Two right images show same silhouettes with noise added to the contour points. First has uniform noise; second has nonuniform noise in patches normal to contour.

the raw test data, we have generated a separate, large, synthetic test set with known pose parameters so that we can obtain error measurements for a variety of experiments on a large volume of data. In order to evaluate our system's robustness to mild changes in the appearance of the object, we generated test sequences in the same manner as the synthetic training set was generated, but with different virtual characters, i.e., different clothing, hair and body proportions. To make the synthetic test set more representative of the real, raw silhouette data, we added noise to the contour point locations. Noise is added uniformly in random directions, or in contiguous regions along the contour in the direction of the 2D surface normal. Such alterations to the contours simulate the real tendency for a simple background subtraction mechanism to produce holes or false extensions along the true contour of the object. (See Figure 4.)

The "true" underlying contours from the clean silhouettes (i.e., the novel silhouettes before their contour points were corrupted) are saved for comparison with the reconstructed silhouettes. The contour error for each frame is defined as the Chamfer distance between the true underlying contours and their reconstructions. For all pixels with a given feature (usually edges, contours, etc.) in the test image $\mathbf{I}$, the Chamfer distance $\mathbf{D}$ measures the average distance to the nearest feature in the template image $\mathbf{T}$.

$$\mathbf{D}(\mathbf{T}, \mathbf{I}) = \frac{1}{N} \sum_{f \in T} d_T(f) \tag{5}$$

where N is the number of pixels in the template where the feature is present, and $d_T(f)$ is the distance between feature f in $\mathbf{T}$ and the closest feature in $\mathbf{I}$.

The pose error for each test frame is defined as the average distance in centimeters between the estimated and true positions of the 19 joints.

Intuitively, a multi-view framework can discern 3D poses that are inherently ambiguous in single-view images. Our results validate this assumption. We performed parallel tests for the same examples, in one case using our existing multi-view framework, and in the other, using the framework outlined above, only with the model altered to be trained and tested with single views alone. Figure 5 compares the error distributions of the single and multi-view

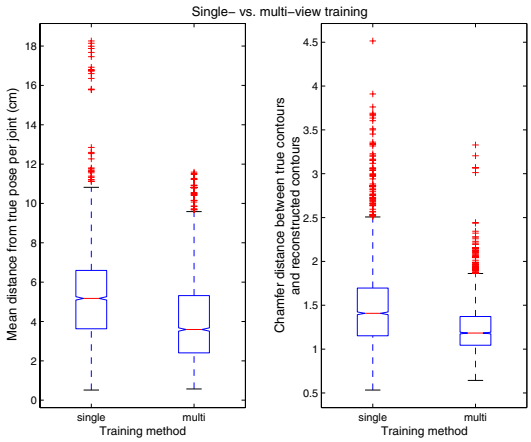


Figure 5: Training on single view vs. training on multiple views. Chart shows error distributions for pose (left) and contours (right). Lines in center of boxes denote median value; top and bottom of boxes denote upper and lower quartile values, respectively. Dashed lines extending from each end of box show extent of rest of the data. Outliers are marked with pluses beyond these lines.

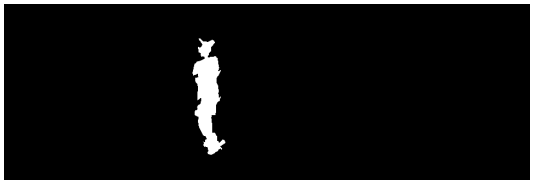
frameworks for a test set of 3,000 examples. Errors in both pose and contours are measured for both types of training. Multi-view estimates are consistently more accurate than single-view estimates. Training the model on multi-view images yields on average 24% better pose inference performance and 16% better contour reconstruction performance than training the model on single-view images.

We have also tested the performance of our multi-view method applied to body pose estimation when only a subset of views is available. A missing view in the shape vector is represented by zeros in the elements corresponding to that view's resampled contour. Just as unknown 3D locations are inferred for the test images, our method reconstructs the missing contours by inferring the shape seen in that view based on examples where all views are known. (See Figures 6, 7, 8, and 9.)

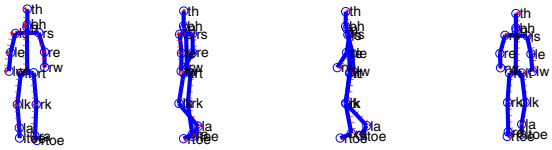
We are interested in knowing how pose estimation performance degrades with each additional missing view, since this will determine how many cameras are necessary for suitable pose estimation should we desire to use fewer cameras than were used in the training set. Once the multi-view model has been learned, it may be used with fewer cameras, assuming that the angle of inclination of the cameras with the ground plane matches that of the cameras with which the model was trained.

Figure 8 shows results for 3,000 examples that have been tested using all possible numbers of views (1,2,3,4), alternately. For a single missing view, each view is omitted systematically one at a time, making 12,000 total tests. For two or three missing views, omitted views are chosen at random

(a) Actual silhouettes (withheld)

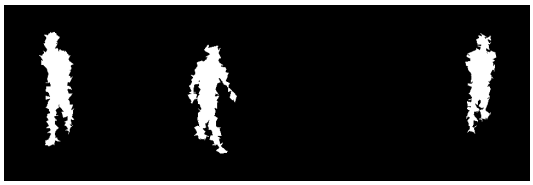


(b) Noisy input

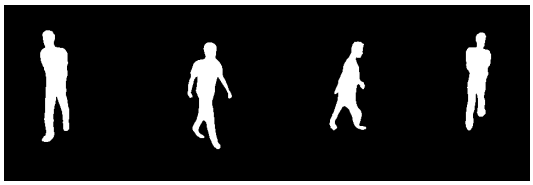


(c) Output

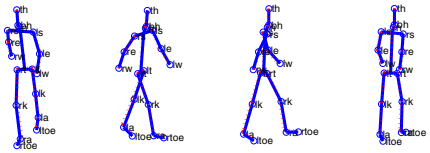
Figure 6: Pose inference from only a single view. Top row shows ground truth silhouettes that are not in the training set. Noise is added to the contour points of second view (middle), and this single view alone is matched to the multi-view shape model in order to infer the 3D joint locations (bottom, solid lines) and compare to ground truth (dotted lines). Abbreviated body part names appear by each joint. This is an example with average pose error of 5 cm.



(a) Noisy input



(b) Reconstructed contours



(c) Output

Figure 7: Pose inference and contour reconstruction with one missing view. Top row shows noisy input silhouettes, middle row shows reconstructed contours (including the inferred shape of the third view), and bottom row shows inferred 3D joint locations (solid lines) and ground truth (dotted lines). This is an example with average pose error of 2.5 cm and a Chamfer distance from the true clean silhouettes of 2.3.

in order to approximately represent all possible combinations of missing views equally. As the number of missing views increases, performance degrades more gracefully for pose inference than for contour reconstruction.

To interpret the contour error results in Figures 5 and 8, consider that the average contour length is 850 pixels, and the silhouettes have an average area of 30,000 pixels. If we estimate the normalized error to be the ratio of average pixel distance errors (number of contour pixels multiplied by Chamfer distance) to the area of the figure, then a mean Chamfer distance of 1 represents an approximate overall error of 2.8%, distances of 4 correspond to 11%, etc. Given the large degree of segmentation errors imposed on the test sets, these are acceptable contour errors in the reconstructions, especially since the 3D pose estimates (our end goal) do not suffer proportionally.

The pose error results are given in real world distances (cm) and are thus fairly straightforward to interpret. Chance performance for pose estimation (i.e., a random draw of pose parameters from the training set) would yield a mean pose error of 29 cm; using four views our method achieves a mean pose error of only 3 cm.

Finally, we evaluated our algorithm on a large data set of real images of pedestrians taken from a database of 4,000 real multi-view frames. The real camera array is mounted on the ceiling of an indoor lab environment. The external parameters of this real four-camera system are roughly the same as those of the virtual cameras in the graphics software that were used for training. The data contains 27 different pedestrian subjects.

Sample results for the real test data set are shown in Figure 9. The original textured images, the extracted silhouettes, and the inferred 3D pose are shown. Without having point-wise ground truth for the 3D locations of the body parts, we can best assess the accuracy of the inferred pose by comparing the 3D stick figures to the original textured images. To aid in inspection, the 3D stick figures are rendered from manually selected viewpoints so that they are approximately aligned with the textured images.

In summary, our experiments show that a probabilistic

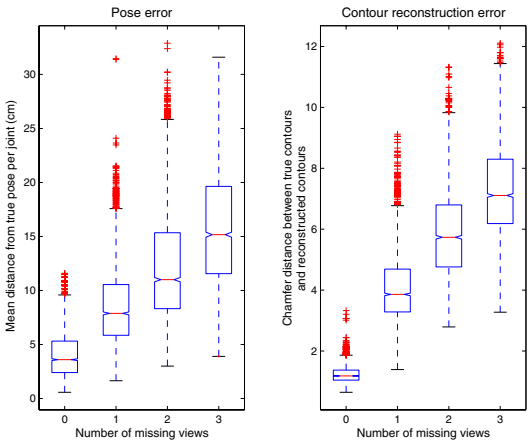


Figure 8: Missing view results. Chart shows error distributions for pose (left) and contours (right) when model is trained on four views, but only a subset of views is used as input to our algorithm. Plotted as in Figure 5.

shape+structure model is able to directly infer 3D structure from observed multi-view image features. Our tests with a large set of noisy, ground-truthed synthetic images offer evidence of our method's ability to infer 3D parameters from contours, even when inputs have segmentation errors. As shown in Figure 8, structure inference for body pose estimation is accurate within 3 cm on average. Performance is good even when there are fewer views available than were used during training; with only one input view, pose is still accurate within 15 cm on average, and can be as accurate as within 4 cm. Finally, we have successfully applied our synthetically-trained model to real data and a number of different subjects.

7. Conclusions and Future Work

We have developed an image-based approach to infer 3D structure parameters using a probabilistic multi-view shape model. Novel examples with contour information but unknown 3D point locations are matched to the model in order to infer the unknown parameters. All computation is done entirely in the image domain and requires no explicit 3D construction. A class-specific prior on multi-view imagery enables accurate estimation of structure parameters in spite of large segmentation errors or even missing input views.

In future work we will explore non-parametric density models, and we will run experiments using motion capture data so that we may compare real image results to ground-truth joint angles. We also intend to include dynamics to strengthen our model for the pedestrian walking sequences. Finally, we are interested in applying our technique in a higher-level gesture or gait recognition system.

References

[1] A. Baumberg and D. Hogg. Learning Flexible Models from Image Sequences. In *Proceedings of European Conference on Computer Vision*, Stockholm, Sweden, May 1994.

[2] A. Baumberg and D. Hogg. An Adaptive Eigenshape Model. In *British Machine Vision Conference*, pages 87–96, Birmingham, England, Sept 1995.

[3] T. Cootes and C. Taylor. A Mixture Model for Representing Shape Variation. In *British Machine Vision Conference*, pages 110–119, Essex, England, 1997.

[4] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan 1995.

[5] T. Cootes, G. Wheeler, K. Walker, and C. Taylor. View-Based Active Appearance Models. *Image and Vision Computing*, 20:657–664, 2002.

[6] M. Covell. Eigen-Points: Control-Point Location Using Principal Component Analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 122–127, Killington, VT, Oct 1996.

[7] Egisys Co. Curious Labs. Poser 5 : The Ultimate 3D Character Solution. 2002.

[8] K. Grauman, G. Shakhnarovich, and T. Darrell. A Bayesian Approach to Image-Based Visual Hull Reconstruction. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Madison, WI, June 2003.

[9] J. Haslam, C. Taylor, and T. Cootes. A Probabilistic Fitness Measure for Deformable Template Models. In *British Machine Vision Conference*, pages 33–42, York, England, Sept 1994.

[10] M. Jones and T. Poggio. Multidimensional Morphable Models. In *Proceedings of the International Conference on Computer Vision*, pages 683–688, Bombay, India, January 1998.

[11] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, Feb 1994.

[12] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-Based Visual Hulls. In *Proceedings ACM Conference on Computer Graphics and Interactive Techniques*, pages 369–374, 2000.

[13] B. Moghaddam. Principal Manifolds and Probabilistic Subspaces for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):780–788, June 2002.

[14] E-J. Ong and S. Gong. The Dynamics of Linear Combinations. *Image and Vision Computing*, 20(5–6):397–414, 2002.

[15] M. Tipping and C. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 1999.

[16] M. Turk and A. Pentland. Face Recognition Using Eigenfaces. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–590, Maui, HI, June 1991.

[17] Y. Wang and L. H. Staib. Boundary Finding with Prior Shape and Smoothness Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):738–743, 2000.

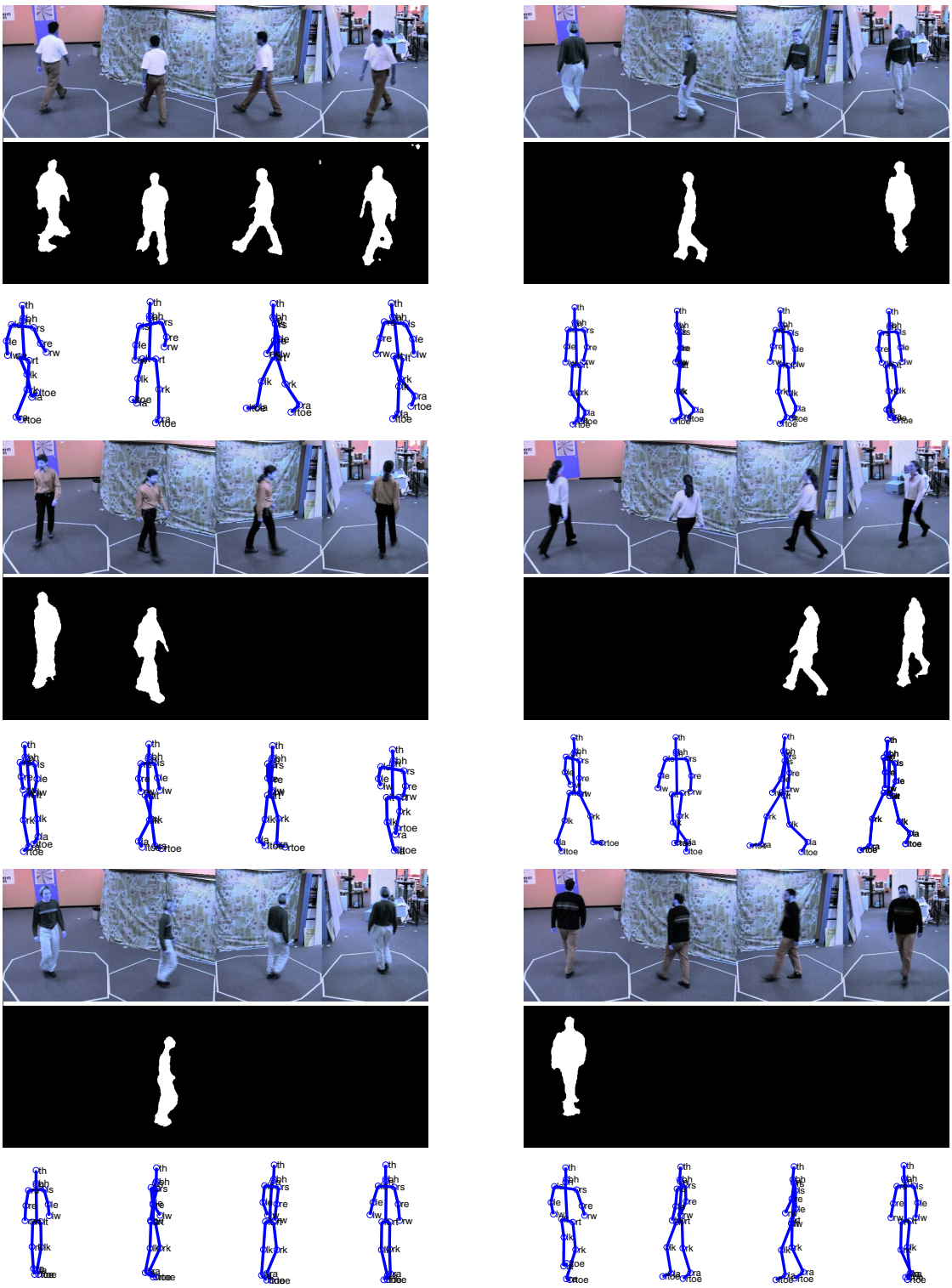


Figure 9: Inferring structure on real data. For each example, top row shows original textured multi-view image, middle row shows extracted input silhouettes where the views that are not used to infer pose parameters are omitted, and bottom row shows inferred joint locations. To aid in inspection, the 3D stick figures are rendered from manually selected viewpoints chosen so that they are approximately aligned with the textured images. In general, estimation is accurate and agrees with the perceived body configuration.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE