Monocular Human Motion Capture with a Mixture of Regressors

Ankur Agarwal and Bill Triggs
GRAVIR-INRIA-CNRS, 655 Avenue de l'Europe, Montbonnot 38330, France
{Ankur.Agarwal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

Abstract

We address 3D human motion capture from monocular images, taking a learning based approach to construct a probabilistic pose estimation model from a set of labelled human silhouettes. To compensate for ambiguities in the pose reconstruction problem, our model explicitly calculates several possible pose hypotheses. It uses locality on a manifold in the input space and connectivity in the output space to identify regions of multi-valuedness in the mapping from silhouette to 3D pose. This information is used to fit a mixture of regressors on the input manifold, giving us a global model capable of predicting the possible poses with corresponding probabilities. These are then used in a dynamical-model based tracker that automatically detects tracking failures and re-initializes in a probabilistically correct manner. The system is trained on conventional motion capture data, using both the corresponding real human silhouettes and silhouettes synthesized artificially from several different models for improved robustness to inter-person variations. Static pose estimation is illustrated on a variety of silhouettes. The robustness of the method is demonstrated by tracking on a real image sequence requiring multiple automatic re-initializations.

1. Introduction

We consider the problem of deducing the 3D pose or motion of the complete human body from a single image or a monocular sequence of images. With potential applications including human computer interaction, image based motion capture, automated visual surveillance and activity recognition in videos/images, this largely unsolved problem has attracted a lot of attention in the computer vision community. Although the estimation and tracking of 3D configurations of complex articulated objects is also studied in the domains of hand and upper body gesture recognition, and there is a large overlap in the technical issues addressed in these areas, the complete human body case poses additional difficulties owing to the large number of degrees of freedom (~ 30 d.o.f. to capture the basic articulations), variations among people and the large range of possible camera orientations. Many of these difficulties are also responsible for the fact that vision-based (markerless) motion capture systems have not currently reached the accuracy of commercially available marker-based ones. However, such systems would provide an easy-to-use, low-cost alternative that would prove useful for tasks such as gesture recognition for human computer interaction where inaccuracies of a few centimeters may be tolerable.

In this paper, we consider the case of a single person, segmentable from the background. We take a learning based approach that statistically models the mapping from 2D human silhouettes to 3D pose configurations using data generated from optical sensor based motion capture. Several forms of learning based methods have recently been explored in this domain, including example or nearest-neighbour based methods [11, 16, 22] and ones that attempt to learn smooth transformations between appearance features and pose space [1, 3, 6]. Many of these methods reduce image observations to silhouette shapes in order to obtain clear input signals. However, one of the key problems with such representations — the occurrence of multiple possible reconstructions from a single image — has yet to be addressed cleanly. Depending on the placement of the camera and the pose of the person, inferring a unique pose from a silhouette often proves very difficult as the mapping from image observations to 3D pose is *multi-valued*. This causes interpolation and function based methods that assume a single-valued mapping from image to pose to fail occasionally, as they are forced to choose or compromise among several possible pose solutions. In this paper, we propose a method that explicitly calculates several possible pose hypotheses from a single silhouette.

To encode silhouette shape, we use descriptors based on histograms of shape contexts [4] as described in [1]. This method encodes silhouette shapes robustly as 100-D descriptors $\mathbf{z}$. In our approach, these observation vectors $\mathbf{z}$ are then reduced to a feature space $\phi(\mathbf{z})$ by performing kernel-PCA [15]. We think of this as a manifold within the original 100-D silhouette space, which folds over onto itself in regions where the silhouettes corresponding to several different body configurations are indistinguishable. During the learning phase, we locally disambiguate these regions using connectivity information on a local neighborhood graph in the pose space $\mathbf{x}$. These local *clusters* are then used as a

starting point to learn a mixture of regressors over the entire space. The mixture retains the local structure that resolves the ambiguities, while at the same time capturing the global relationship between $\mathbf{z}$ and $\mathbf{x}$. The result is a set of nonlinear regressors that is capable of predicting the different possible pose solutions with their corresponding probabilities of occurrence.

These multivalued pose estimates are very valuable when tracking motions through a sequence of images because they allow the system to deal directly with multiple hypotheses. Tracking systems that rely mainly on temporal continuity are susceptible to tracking failures, and including detection at each frame can help to detect and recover from such failures (*e.g.* [24]). In our case, our probabilistic pose estimator provides both an observation density and detection of loss of track, allowing the tracker to automatically reinitialize whenever necessary *e.g.* if the subject reappears after moving out of the field of view.

To train our system, we supplement a database of real human silhouettes obtained from motion capture videos with a set of silhouettes synthesized artificially from several different human models using the motion capture poses. (See figure 1.)

Figure 1. Some sample silhouettes from our synthetic training database, synthesized from motion capture data. The database consists of 16,000 images: 8 different characters (2 of which are reserved for test purposes) in poses from 8 motion sequences, capturing inter-person variations, and typical walking movements plus some casual gestures. Note that with many silhouettes, it is hard to tell right and left limbs, or front and back-facing poses apart.

Previous work: There is a large literature on human pose estimation and tracking in images. Most of the existing approaches use multiple cameras to capture 3D depth information, but here we restrict ourselves to the monocular case owing to its wide applicability. A variety of methods based on explicit human body models, geometry and optimization have been shown to be effective [17, 20, 13]. Another class of methods avoids the need to build hand-tailored body models and cost functions by learning from a database of training examples, the main goal being to construct a model that provides efficient generalization to new examples. Some of these approaches (*e.g.* [16]) use essentially local methods (typically based on nearest neighbour search), while others attempt to exploit more global structure in the relationship between appearance and pose (*e.g.* [1, 5, 6]), possibly with an implicit encoding of locality based on the use of a kernel. The main difficulty with all of these methods is the fact that the pose-to-silhouette mapping is many-to-many: different-looking silhouettes often have the same pose, and a given silhouette can often arise from several poses. The first issue has been addressed to some extent, *e.g.* by separating *style* from *content* [6], and by developing robust shape descriptors or matching algorithms (*e.g.* [1, 7]), but very little work is available on resolving the one-to-many issue in going from silhouette to pose. The only work that we are aware of that directly addresses this issue is that described in [14] — this uses 'specialized maps' to learn the mapping in the form of several sub-functions, but fails to demonstrate an ability to deal with ambiguities in a probabilistic manner.

A few recent papers have made use of manifold embedding techniques to model silhouette to pose mappings. The authors of [6] use temporal information as a neighbourhood criterion to learn 'activity manifolds'. In [3], clustering in the silhouette space is used to learn multiple mappings, for the much less ambiguous 2D pose estimation problem. The model proposed in this paper uses clustering on a manifold to resolve 3D ambiguities, which allows the reconstructions to be used in an automated tracking environment. This new approach thus attempts to combine tracking methods that are based on dynamics-driven sample propagation and image likelihood measurements (*e.g.* [17, 21]), with more *independent detection* style approaches as in [23, 18]. *c.f.* [2].

An approach somewhat similar to ours, developed indenpendently and contemporaneously, is described in [19].

Organization: The next two sections discuss a mixture-of-Gaussians based scheme for computing a multimodal distribution over 3D poses. Section 4 contains results on probabilistic pose estimation from single images and section 5 describes a self-(re)initializing 3D human tracker based on these probabilistic pose estimates. Section 6 concludes with a discussion and some perspectives.

2. Multimodal Pose Estimation

To account for the missing information in a silhouette, we introduce a latent variable $\mathbf{l}$ that will implicitly capture the limb labelling and kinematic-flipping [21] possibilities of the

given silhouette. The 3D pose $\mathbf{x}$ is encoded as a vector of joint angles. The central assumption is that given the value $\mathbf{l}$, the 3D pose has a functional dependence on the observed silhouette:

$$\mathbf{x} \,|\, \mathbf{l} \;\simeq\; \mathbf{r}_\mathbf{l}(\mathbf{z}) + \boldsymbol{\epsilon}_\mathbf{l} \tag{1}$$

where $\mathbf{z}$ is the observation (silhouette shape descriptor vector), $\mathbf{r}_\mathbf{l}$ is a functional transformation from $\mathbf{z}$ to $\mathbf{x}$, and $\boldsymbol{\epsilon}_\mathbf{l}$ is a noise vector. Modelling $\boldsymbol{\epsilon}_\mathbf{l}$ as a Gaussian with zero mean and covariance $\boldsymbol{\Lambda}_\mathbf{l}$, the conditional pose distribution $p(\mathbf{x} \,|\, \mathbf{z}, \mathbf{l})$ is $\mathcal{N}(\mathbf{r}_\mathbf{l}(\mathbf{z}), \boldsymbol{\Lambda}_\mathbf{l})$.

Latent variables in the form of explicit left/right limb labellings have been used in some areas (*e.g.* [8]). This could be extended to labelling the 3D kinematic flipping possibilities (motions towards/away from the camera that leave the image unchanged) that represent the main residual reconstruction ambiguity once the image limbs have been labelled. However, this would require an exponential number of labels to account for all of the flipping possibilities across all limbs. In practice, such a fine level of labelling is not really needed to disambiguate between the probable pose hypotheses. We find that there are typically only a hand-full of probable *modes* (3D pose reconstructions). One the other hand, the multiplicity of solutions usually persists over considerable subspaces within the silhouette space. We thus choose not to attach an explicit meaning to our latent variables but rather learn their values automatically for different silhouettes so as to capture the essence of the information required to disambiguate between typical human poses. We model our latent variable as belonging to a discrete set[1]: $\mathbf{l} \in \{1, 2 \ldots K\}$.

Marginalizing over all possible values of the latent variable for a given observation $\mathbf{z}$, we obtain

$$p(\mathbf{x} \,|\, \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{l}{=}k \,|\, \mathbf{z}).\, \mathcal{N}(\mathbf{r}_k(\mathbf{z}), \boldsymbol{\Lambda}_k) \tag{2}$$

The pose estimator thus takes the form of a mixture of uncertain regressors.

3. Mixture of Regressors

Our silhouette representation is redundant and silhouette-to-pose regression is highly nonlinear, so it turns out to be useful to perform a kernel PCA to extract relevent dimensions of the input feature space, giving us a reduced representation $\boldsymbol{\phi}(\mathbf{z})$. This can be imagined as lying on a manifold within the silhouette descriptor space that is folded over onto itself due to many-to-one projection mappings. To allow for multimodal output distributions, the mapping to the output space is then learned as a mixture of regressors (often known as a mixture of experts [10]) on $\boldsymbol{\phi}(\mathbf{z})$. In practice, we use a polynomial dot product kernel based on the Bhattacharya measure for histogram similarity: $K(\mathbf{z}_1, \mathbf{z}_2) = \langle \sqrt{\mathbf{z}_1}, \sqrt{\mathbf{z}_2} \rangle^p$. With p=6, we reduce 100-D vectors $\mathbf{z}$ to 23-D vectors $\boldsymbol{\phi}(\mathbf{z})$. This is found to be much more effective in practice than, *e.g.*, performing a linear PCA, which requires 91 dimensions to retain 99% variance in our data.

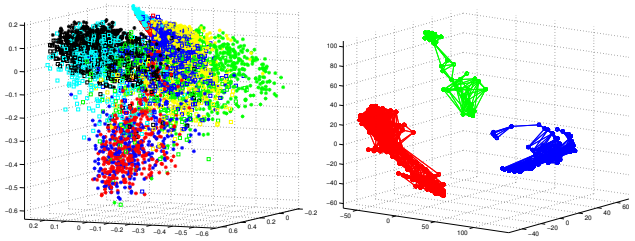

Figure 2. *(Left)*: Initial clusters in $\boldsymbol{\phi}(\mathbf{z})$ obtained by running k-means with k=12. A projection on the first 3 kernel principal components is shown. *(Right)*: 3 connected components are obtained for one of these clusters, as seen on the neighbourhood graph of the corresponding points in $\mathbf{x}$. This cluster is thus split into 3 to separate the different pose subclasses that it contains. Of the 12 initial clusters in $\boldsymbol{\phi}(\mathbf{z})$, we find that 3 get split into 2 components each and 2 into 3 components each based on this connectivity analysis. A few of these merge into others during the EM process, giving a final model consisting $\sim$20 clusters.

3.1. Connectivity based clustering

The key to successful learning is to clearly separate the ambiguous cases into different mixture components (clusters) at initialization. Otherwise the individual regressors tend to average over several possible solutions. For this, we first use k-means to divide the $\mathbf{z}$ space into several clusters in the KPCA-reduced space $\boldsymbol{\phi}(\mathbf{z})$. (This corresponds to performing a spectral clustering in the original space $\mathbf{z}$ [12].) Each of these clusters is then split into subclusters by making use of the corresponding $\mathbf{x}$ values, exploiting the fact that silhouettes that appear similar in $\boldsymbol{\phi}(\mathbf{z})$ can be disambiguated based on the distance between their corresponding 3D poses. For this, we construct a neighbourhood graph in $\mathbf{x}$ (which we assume to encode the *true* distance between points), having an edge between all points within a thresholded distance from one another, and robustly identify connected components in this graph for each cluster in $\boldsymbol{\phi}(\mathbf{z})$. The scheme is illustrated in figure 2. We find that this two-step clustering gives better performance than the other initialization methods that we tested. For example, in terms of final reconstruction errors on a test set after EM based learning (see below), clustering in either $\mathbf{x}$ alone or jointly in $(\mathbf{x}, \boldsymbol{\phi}(\mathbf{z}))$ is found to give re-

[1]Besides the small number of typically possible reconstructions, other attributes that can potentially be captured by latent variables — such as inter-person variations — can also be thought of as being discrete, *e.g.* a finite number of 'person classes'. In practice, though, we find that a robust silhouette representation is already insensitive to many of these inter-person variabilities.

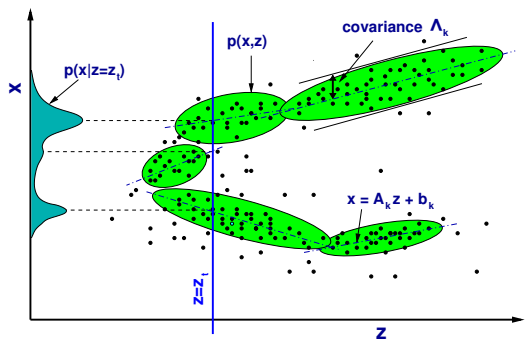

Figure 3. An illustration of the density estimation / regression model used to estimate the conditional density $p(\mathbf{x} \mid \mathbf{z})$.

construction errors higher by 0.3 degrees on average, while clustering in $\boldsymbol{\phi}(\mathbf{z})$ alone shows several instances of averaging across multiple solutions owing to the inability to resolve the ambiguities, also increasing the average error.

3.2. Expectation-Maximization based learning

The clusters obtained above are used to learn a global probabilistic model by fitting a mixture of Gaussians to the joint density of $(\boldsymbol{\phi}(\mathbf{z}), \mathbf{x})$:

$$\begin{pmatrix} \boldsymbol{\phi}(\mathbf{z}) \\ \mathbf{x} \end{pmatrix} \simeq \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k) \tag{3}$$

where π_k are the gating probabilities $p(\mathbf{l} = k)$ of the respective classes. Figure 3 shows the different components of the model. This model is estimated using EM, but given that our goal is to estimate a regressive model (the conditional density $p(\mathbf{x} \mid \boldsymbol{\phi}(\mathbf{z}))$), not a generative one (the joint density $p(\boldsymbol{\phi}(\mathbf{z}), \mathbf{x})$) we impose a particular covariance structure within the EM framework. Within each class, the regressor is modelled as

$$\mathbf{x} = \mathbf{r}_k(\mathbf{z}) + \boldsymbol{\epsilon}_k \;\equiv\; \mathbf{A}_k \, \boldsymbol{\phi}(\mathbf{z}) + \mathbf{b}_k + \boldsymbol{\epsilon}_k \tag{4}$$

where $\boldsymbol{\epsilon}_k$ has a constant covariance $\boldsymbol{\Lambda}_k$ independent of $\mathbf{z}$. This implies the following relations for the mean and covariance of each class:

$$\boldsymbol{\mu}_k = \begin{pmatrix} \boldsymbol{\phi}(\bar{\mathbf{z}}_k) \\ \mathbf{r}_k(\bar{\mathbf{z}}_k) \end{pmatrix}, \boldsymbol{\Gamma}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k & \boldsymbol{\Sigma}_k \mathbf{A}_k^\top \\ \mathbf{A}_k \boldsymbol{\Sigma}_k & \mathbf{A}_k \boldsymbol{\Sigma}_k \mathbf{A}_k^\top + \boldsymbol{\Lambda}_k \end{pmatrix} \tag{5}$$

This model has conditional covariance of $\boldsymbol{\Lambda}_k$ for $\mathbf{x} \mid \mathbf{z}$ (the vertical "thickness" of the classes in figure 3), to which is added the uncertainty $\mathbf{A}_k \boldsymbol{\Sigma}_k \mathbf{A}_k^\top$ inherited from $\mathbf{z}$ via $\mathbf{A}$ to form the full covariance for $\mathbf{x}$. To avoid overfitting, we assume the descriptor covariance matrices $\boldsymbol{\Sigma}_k$ and the residual noise covariances $\boldsymbol{\Lambda}_k$ to be diagonal. Our 'M' step now consists of two parts: First, $\mathbf{A}_k, \mathbf{b}_k$ are estimated by weighted least squares regression (each example being weighted by its responsibility for the given class) using the linear model given in (4), followed by estimating the covariances $\boldsymbol{\Lambda}_k$ from the residual errors. Second, the statistics $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ for each class are computed, given the class membership weights for each point (initialized from the clustering above). The 'E' step, as usual, involves re-estimating the membership weights (responsibilities) for each point given the statistics of each class. The process is iterated to convergence, which takes 30-40 iterations. Occasionally, a few of the clusters 'die out' as their points are merged with others.

The EM process 'smooths' the initial clusters, giving better generalization in terms of test set performance by exploiting the global mapping between the $\mathbf{z}$ and $\mathbf{x}$ spaces. At the same time, the initial *structure* contained in the connectivity based clustering is retained, as is seen by observing the final clusters.

For inference, we are given $\mathbf{z}$ and we need to estimate $\mathbf{x}$. For this we use (2). The conditional probabilities $p(\mathbf{l}{=}k \mid \mathbf{z})$ can be computed from the observations alone:

$$p(\mathbf{l}{=}k \mid \mathbf{z}) \;=\; \frac{\pi_k \cdot \mathcal{N}(\boldsymbol{\phi}(\bar{\mathbf{z}}_k), \boldsymbol{\Sigma}_k) \mid_{\boldsymbol{\phi}(\mathbf{z})}}{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\boldsymbol{\phi}(\bar{\mathbf{z}}_k), \boldsymbol{\Sigma}_k) \mid_{\boldsymbol{\phi}(\mathbf{z})}} \tag{6}$$

where $\mathcal{N}(\boldsymbol{\phi}(\bar{\mathbf{z}}_k), \boldsymbol{\Sigma}_k) \mid_{\boldsymbol{\phi}(\mathbf{z})}$ is the normal function with mean $\boldsymbol{\phi}(\bar{\mathbf{z}}_k)$ and covariance $\boldsymbol{\Sigma}_k$, evaluated at the point $\boldsymbol{\phi}(\mathbf{z})$.

	% of frames with m solutions			Error in the top solution	Error in best of top 4 solutions
	$m = 1$	$m = 2$	$m \geq 3$		
(A)	62	28	10	6.14	4.84
(B)	65	28	6	7.40	5.37
(C)	72	23	5	6.14	4.55

Figure 4. The numbers of solutions and the errors (RMS of joint angles in degrees) obtained when reconstructing three different datasets. To count the number of modes predicted, we consider only modes with $p > 0.1$.

4. Experimental Performance

In this section we analyze the accuracy of our mixture model for the estimation of full 3D body pose from single silhouette images. We represent body pose as 54 joint-angle values to correspond to the motion capture data. The system is trained on 8000 silhouettes from the database illustrated in figure 1 and tested on images of real people.

To quantify performance, we first measure accuracy on two test cases involving silhouettes artificially synthesized using poses from human motion capture data: a test set (A) consisting of ∼600 frames of a person not included in the training data, and a test set (B) consisting of ∼400 frames of a person in the training data but with a different motion sequence. For comparison, we also report errors in a subset (C) of ∼600 frames from the original training set. We find that

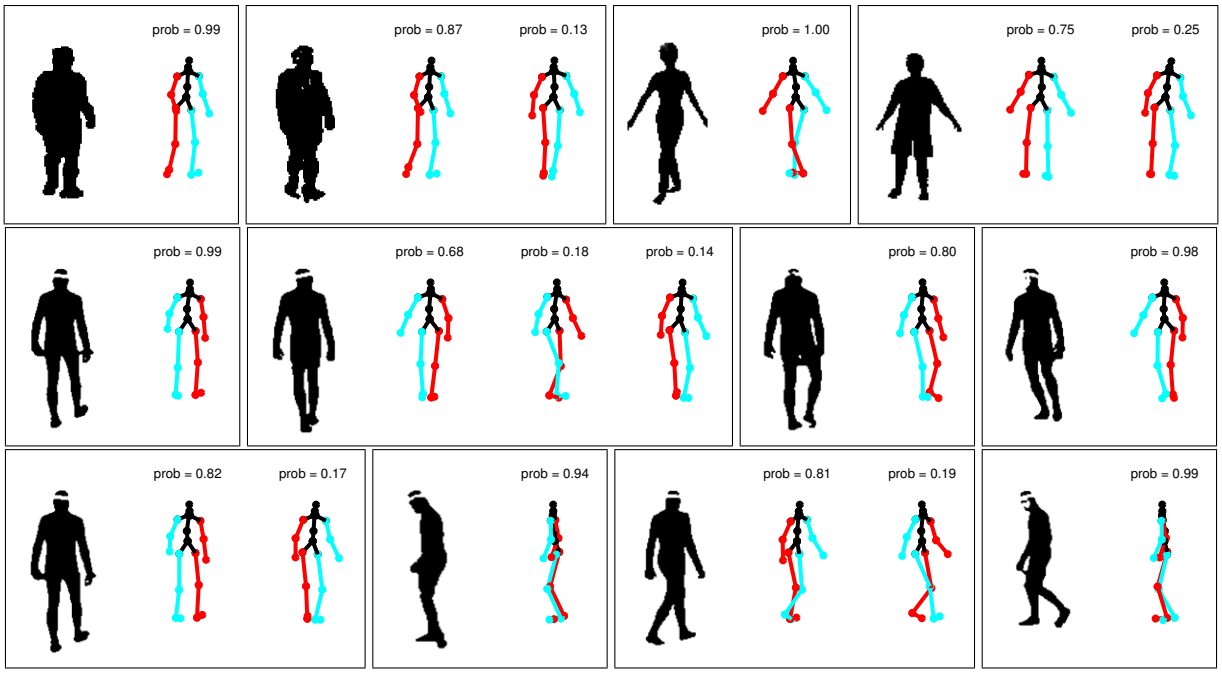

Figure 5. 3D pose estimates obtained on test silhouettes of real people not present in the training set. The mean pose estimates from all modes with probability greater than 0.1 are displayed, with red / dark gray denoting right and blue / light gray denoting left limbs. The method mostly gives accurate pose reconstructions, identifying the possible solutions well when there are ambiguities. Some interesting and non-obvious forward-backward ambiguities in the silhouettes are revealed. *E.g.* in the second-last reconstruction, the two cases correspond to the person walking *into* the image plane, 45° towards the right, or *out* of the image plane, again 45° to the right. Note that the arms are clearly interchanged in the 2 cases.

the mixture generally outputs between 1 and 3 high probability solutions for each silhouette, with the highest probability solution often but not always being the correct one. When errors occur, it is typically also hard for a human to decide between the different reconstructions. The statistics are summarized in figure 4. We also plot the accuracy of the most probable regressor, and the accuracy of the best (closest to the ground truth) of the four most likely regressors in terms of errors computed as average RMS deviations for each joint-angle. The better overall performance on test (A) than on test (B) suggests that the model generalizes better between different appearances than between different motion patterns.

Figure 5 shows sample reconstructions on some real test silhouettes, none of which were included in the training data. Since ground truths were not available for all of these silhouettes, we visually inspected the reconstructions on 300 frames of one of the real test sequences to quantify the quality of reconstructed modes. We find that in 47% of images, the highest ranked solution gives a good reconstruction. In respectively 24% and 13% of the images, the second / the third or fourth modes give suitable reconstructions, while in the remaining 16%, the system fails to give the correct pose estimate within the top 4 modes. One place where this happens is at the points where multiple surfaces split or merge in the $\mathbf{z}-\mathbf{x}$ space. As seen in the last example in the figure, the regressors sometimes still average over multiple poses (the two possible leg labellings) that are very similar in $\mathbf{z}$, and that are also not too far apart in $\mathbf{x}$. The local connectivity information used in the learning process sometimes fails to disambiguate such regions as the ambiguous solutions are very close to each other in pose space, causing them to belong to the same connected region in the pose neighbourhood graph.

We also tested the effect of training on different people. The generalization to new people improves when more people are added to the training database. However, as shown in figure 6, the improvement is relatively minor as regards the accuracy of individual regressors, although the ability to select the correct solutions shows a slightly more significant improvement. Overall the method generalizes among people surprisingly well. We attribute this to the robust representation of the silhouette shape [1], the descriptor feature space having been learned on several people.

5. Self-Initialized 3D Tracking

Our multiple hypothesis pose estimator can also be used to give smooth reconstructions of 3D human body motions in a

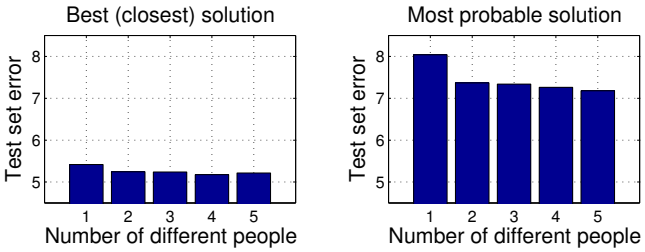

Figure 6. Generalization performance across different people in our 8 character dataset. By varying the number of people in the training set from one to five and testing on all of the others, we find that there is very little improvement in the performance of individual regressors (as seen by the accuracy of the closest (to ground truth) solution, but a slightly more significant increase in the system's ability to select probable solutions. The method appears to generalize between different people quite well. The numbers were obtained using k-fold cross validation in each case.

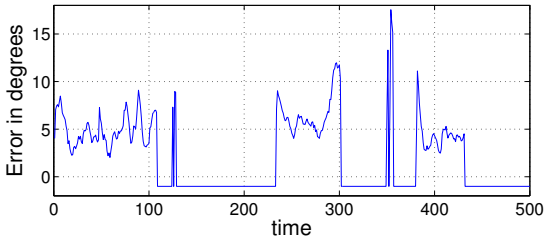

Figure 8. Error of the tracker (RMS deviations for individual joint angles) across the sequence shown in figure 7. A value of -1 indicates that no person is detected and the tracker is turned 'off'.

video sequence. We demonstrate this using a Condensation [9] based tracker initialized robustly with our multimodal pose estimate.

We use a stochastic first order autoregressive process to model the dynamics from the motion capture data, and assign likelihoods to the state particles by treating our multimodal pose estimates as *observation densities* in pose space. Besides providing the ability to automatically initialize from the mixture of Gaussian density modelled on $\phi(\mathbf{z})$, our pose estimator provides estimates of the probability that a valid human body (shape) observation was seen. These are used to detect images where the person is not observed and to accordingly re-initialize when he/she becomes visible again, allowing tracking to continue after failures. We currently use a simple threshold on 'human presence probability' in order to re-initialize particles from the current observation. A probabilistic re-initialization scheme is under development. Re-initialization is also triggered when the total particle density falls to an unexpectedly low level, indicating a failure in tracking for some other reason.

Figure 7 shows sample frames from the tracking of a real motion sequence in which the subject disappears from the field of view several times. The body pose is successfully tracked through the 500 frames, the tracker being automatically (re)initialized at $t = 1$, 234 and 389. Although the initializations are not always perfect, multiple hypothesis tracking allows the correct modes to emerge after a few frames, giving a stable track. The overall error in pose estimation across the sequence is shown in figure 8. Rapid stabilization is seen in both cases of reinitialization at $t = 234$ and 389. Instances of false detection and initialization are visible at $t \sim 125$ and 350. The reconstructions in figure 7 show the most likely particle at any given instant, but do not necessarily reflect the optimal temporal sequence of state estimates. The latter may be obtained, *e.g.*, by backtracing the particles that contributed highly likely tracks — a mechanism that also resolves the ambiguities present in individual images by exploiting temporal coherency.

For this experiment, the model was trained on different sequence of motions by motion capture data and silhouettes taken from the same subject, in order to allow for a broader coverage of his range of motions. This demonstrates the method's applicability to different kinds of data, as we work with real images and standard motion capture formats. However, some technicalities needed to be handled in this respect. The Euler angle based pose representation used by the training data is subject to "gimbal lock" singularities. We currently remove this phenomenon by re-ordering the component rotations, assuming that complete 360 degree rotations occur only along a single axis at each joint. This is reasonable in general but it would prevent smooth tracking in unusual some cases, *e.g.* for a person walking in circles and performing somersaults at the same time. We are currently exploring re-parameterizations based on quaternions and 3D joint locations to resolve this potential problem.

6. Conclusion

We have described a method for multiple hypothesis estimation of 3D human pose from silhouettes, based on mixtures of regressors. This overcomes the problem of ambiguities present in silhouette data. The method combines the advantages of example based methods, exploiting information on a silhouette manifold and pose neighbourhood graph, and global structure in the form of a single probabilistic model. Accurate pose reconstruction results are obtained on a variety of real unseen silhouettes, demonstrating the method's ability to generalize across inter-person variations and imperfect silhouette extraction.

When used in a multiple hypothesis tracker, the method is capable of tracking stably over time with robustness to occasional tracking failures. The state density is automatically

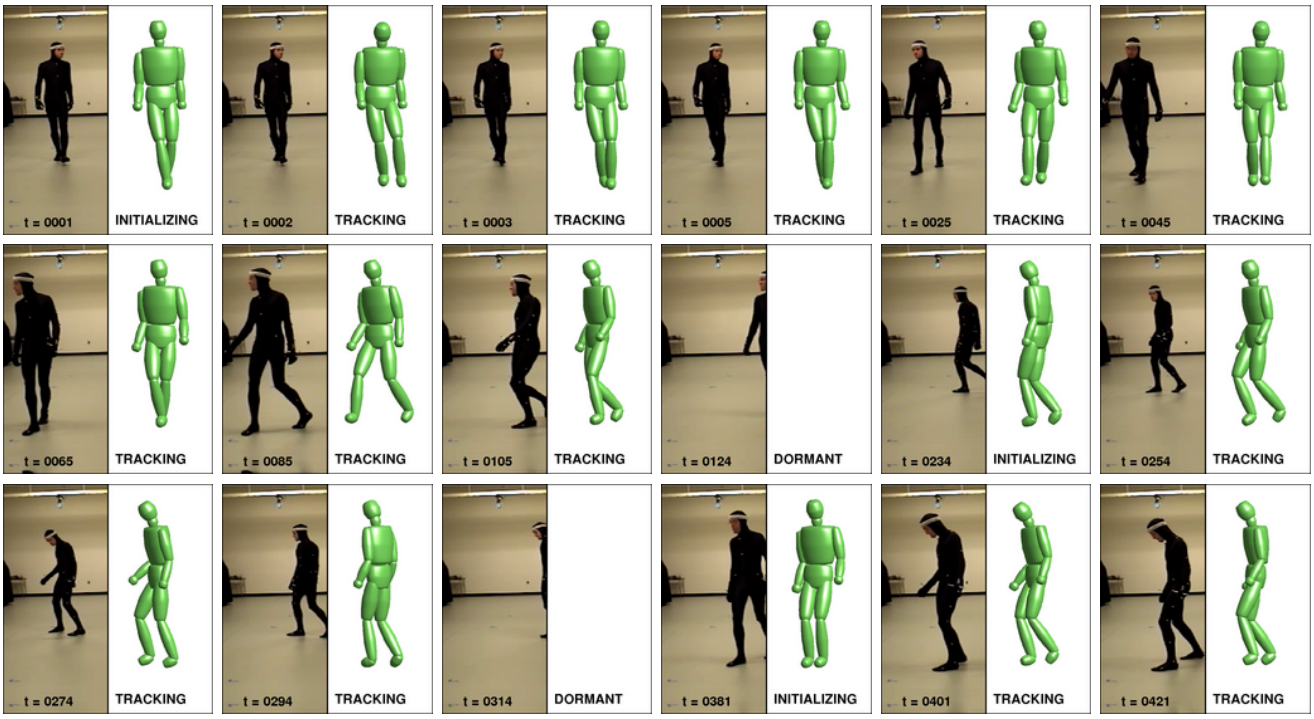

Figure 7. Snapshots from multiple hypothesis tracking of a person across 500 frames. (Sequence taken from http://mocap.cs.cmu.edu/.) Our *direct* and probabilistic pose estimation from the image allows automatic initialization, and re-initialization on detecting tracking failure or absence of a person (see text). Maintaining multiple track hypotheses allows the tracker to recover from possibly inaccurate initializations, tracking stably through instances where the person is not observed. The overall error with time for this track is shown in figure 8.

reinitialized on detecting a person again after periods of non-observation.

Ongoing and future work: We are currently improving the tracker to be fully probabilistic with respect to detecting tracking failures and reinitializing. We would also like to generalize it to tracking in more complex environments, where segmentation based on background subtraction is not possible. Some initial experiments on pose estimation from raw unsegmented input images have shown that the method also works to some extent without prior segmentation, but the observation respresentation needs to be remodelled to handle clutter better. Other possible extensions include robustness to occlusions and handling more than one person.

Acknowledgments

This work was carried out under a MENRT Doctoral Fellowship from the French Ministry of National Education, Research and Tecnology; and the European Union project LAVA.

References

[1] A. Agarwal and B. Triggs. 3D Human Pose from Silhouettes by Relevance Vector Regression. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[2] A. Agarwal and B. Triggs. Learning to Track 3D Human Motion from Silhouettes. In *Int. Conf. on Machine Learning*, 2004.

[3] V. Athitsos and S. Sclaroff. Estimating 3D Hand Pose From a Cluttered Image. In *Int. Conf. Computer Vision*, 2003.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition using Shape Contexts. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 24(4):509–522, 2002.

[5] M. Brand. Shadow Puppetry. In *Int. Conf. Computer Vision*, pages 1237–1244, 1999.

[6] A. Elgammal and C. Lee. Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[7] K. Grauman and T. Darrell. Fast Contour Matching Using Approximate Earth Mover's Distance. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[8] I.Haritaoglu, D.Harwood, and L.Davis. Ghost: A Human Body Part Labeling System Using Silhouettes. In *International Conference on Pattern Recognition*, 1998.

[9] M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

[10] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991.

[11] G. Mori and J. Malik. Estimating Human Body Configurations Using Shape Context Matching. In *European Conf. Computer Vision*, volume 3, pages 666–680, 2002.

[12] A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2001.

[13] V. Parameswaram and R. Chellappa. View Independent Body Pose Estimation from a Single Perspective Image. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[14] R. Rosales and S. Sclaroff. Learning Body Pose via Specialized Maps. In *Advances in Neural Information Processing Systems*, 2001.

[15] B. Schölkpof, A. Smola, and K. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

[16] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *Int. Conf. Computer Vision*, 2003.

[17] H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conf. Computer Vision*, volume 1, 2002.

[18] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking Loose-Limbed People. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[19] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Learning to Reconstruct 3D Human Motion from Bayesian Mixture of Experts. In *Int. Conf. Computer Vision & Pattern Recognition*, 2005.

[20] C. Sminchisescu and B. Triggs. Kinematic Jump Processes For Monocular 3D Human Tracking. In *Int. Conf. Computer Vision & Pattern Recognition*, June 2003.

[21] C. Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6):371–391, June 2003. Special issue on Visual Analysis of Human Movement.

[22] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Learning a kinematic prior for tree-based filtering. In *British Machine Vision Conference*, volume 2, pages 589–598, 2003.

[23] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric. *International Journal of Computer Vision*, 48(1):9–19, 2002.

[24] O. Williams, A. Blake, and R. Cipolla. A Sparse Probabilistic Learning Algorithm for Real-Time Tracking. In *Int. Conf. Computer Vision*, 2003.