Discriminative Density Propagation for 3D Human Motion Estimation

Cristian Sminchisescu[1] **Atul Kanaujia**[2] **Zhiguo Li**[2] **Dimitris Metaxas**[2]
[1]Department of Computer Science, University of Toronto, Canada, *crismin@cs.toronto.edu*
[2]Department of Computer Science, Rutgers University, USA, {*kanaujia,zhli,dnm*}*@cs.rutgers.edu*

Abstract

We describe a mixture density propagation algorithm to estimate 3D human motion in monocular video sequences based on observations encoding the appearance of image silhouettes. Our approach is discriminative *rather than generative, therefore it does not require the probabilistic inversion of a predictive observation model. Instead, it uses a large human motion capture data-base and a 3D computer graphics human model in order to synthesize training pairs of typical human configurations together with their realistically rendered 2D silhouettes. These are used to directly* learn *to predict the conditional state distributions required for 3D body pose tracking and thus avoid using the generative 3D model for* inference *(the learned discriminative predictors can also be used, complementary, as importance samplers in order to improve mixing or initialize generative inference algorithms). We aim for probabilistically motivated tracking algorithms and for models that can represent complex multivalued mappings common in inverse, uncertain perception inferences. Our paper has three contributions: (1) we establish the density propagation rules for* discriminative inference *in continuous, temporal chain models; (2) we propose flexible algorithms for* learning *multimodal state distributions based on compact, conditional Bayesian mixture of experts models; and (3) we demonstrate the algorithms empirically on real and motion capture-based test sequences and compare against nearest-neighbor and regression methods.*

Keywords: *density propagation, mixture modeling, hierarchical mixture of experts, 3D human tracking, Bayesian methods, sparse regression.*

1 Introduction and Motivation

We consider the problem of tracking and reconstructing (inferring) 3D articulated human motion in monocular video sequences. This is a challenging research topic with a broad set of applications for scene understanding, but our argument applies more generally to temporal state estimation problems. Approaches to tracking and modeling can be classified as *generative* and *discriminative*. They are similar in that both require a state representation ($\mathbf{x}$) , here a 3D human model with kinematics (*e.g.* joint angles) or shape (surfaces or joint positions), and they both use a set of image features as observations ($\mathbf{r}$) for state inference. (Often, a training set, $\mathcal{T} = \{(\mathbf{r}_i, \mathbf{x}_i) \mid i = 1 \ldots N\}$ sampled from the *joint distribution* is available.) The computational goal for both approaches is common: the conditional distribution, or a point estimate, for the model state, given observations.

Generative algorithms typically model the joint distribution using a constructive form of the the observer: the observation likelihood or cost function. Inference involves complex search over the state space in order to locate the peaks of the likelihood, *e.g.* using non-linear optimization or sampling. Bayes' rule is then used to compute the state conditional from the observation conditional and the state prior. Learning can be both supervised and unsupervised. This includes priors on the state [10, 12, 21], dimensionality reduction [22] or estimating the parameters of the observation model (*e.g.* texture, ridge or edge distributions) using problem-dependent, natural image statistics [19]. Temporal inference (tracking) is framed in a clear probabilistic and computational framework based on mixture or particle filters [13, 10, 25, 21, 26].

It has been argued that generative models can flexibly reconstruct complex unknown motions and can naturally handle problem constraints. It has been counter-argued that both flexibility and modeling difficulties lead to expensive, uncertain inference [10, 25, 23, 21], and that a constructive form of the observer is somewhat indirect with respect to the task, that requires conditional state estimation and not conditional observation modeling.

These arguments motivate the complementary study of **discriminative algorithms** [7, 17, 20, 18, 2] that model and predict the state conditional directly in order to simplify inference. Prediction however involves missing (state) data, unlike learning that is supervised. But learning is also difficult because modeling perceptual data requires adequate representations of highly multimodal distributions.[1] While this implies that, strictly, the inverse mapping from observations to states is multi-valued and cannot be functionally (and globally) approximated, several authors made initial progress by treating it so [20, 4, 17, 28, 2]. Some approaches constructed data structures for fast nearest-neighbor retrieval [20, 4, 28, 17] or learned regression parameters [2]. Inference involved either indexing for the nearest-neighbors of the observation and using their state for locally weighted predictions, direct pre-

[1]This reflects the structure of the problem and not a particular modeling. *E.g.* think of conversations observed from a side, where gestures pointing towards or away from the camera are common. Humans can initiate a large variety of motions starting from passive (*e.g.* stand-up) positions. Many state trajectories will intersect and produce ambiguity in such regions.

diction using the learned regressor parameters [2], or affine reconstruction from joint centers [17].

Among discriminative methods, a notable exception is [18], who clustered their dataset into soft partitions and learned functional approximations (perceptrons) within each. However, clusterwise functional approximation [9, 18] is only going halfway towards a multivalued inversion because inference is not straightforward. For new inputs, cluster membership probabilities cannot be computed as during (supervised) learning, because the state is missing. The joint mixture coefficients are not useful either because they are the fixed cluster membership averages over the training set. Therefore the extent to which each perceptron is good at predicting a given output is not easy to compute. On the other hand, averaging predictions from multiple perceptrons can give unsatisfactory results (see fig. 2 for a discussion). Nevertheless, the method is useful as a proposal mechanism, *e.g.* during generative inference based on quadrature-style Monte-Carlo approximations and indeed this is how it has primarily been used [18]. A related method has been proposed by [11], where a mixture of probabilistic PCA is fitted to the joint distribution of multiview silhouettes and corresponding 3D pose, and reconstruction is based on MAP estimates. In this multi-image setting the state conditional could be unimodal, but conditional computation requires, in principle, application of Bayes' rule and marginalization [24].

To summarize, it has been argued that discriminative models provide fast inference and interpolate flexibly in the trained region. But they can fail on novel inputs, especially if trained using small datasets. Increasing the training set or the complexity of motion inevitably leads to multimodal state conditionals (§3). But learning such distributions is difficult and most exiting methods [20, 28, 11, 2] are unimodal. Finally, discriminative methods lack a clear probabilistic temporal estimation framework that has been so fruitful with generative models [13, 10, 25]. Existing tracking algorithms [28, 2] involve per-frame state inference, possibly using estimates at previous timesteps [28, 2], but do not rely on a proven set of independence assumptions or propagation rules. What distributions should be modeled and how should they be combined for optimal solutions?

The research we present has **three contributions:**

(1) We propose a probabilistic framework and derive the density propagation rules for inference in discriminative, continuous chain models. The key ingredients of this approach are: *(a)* the structure of the model (see fig. 1 and §2.1); *(b)* the representation of local, per-node conditional state distributions (see **(2)** below and §2.2); *(c)* the belief propagation (chain inference) procedure (§2.1). Here we work parametrically (and analytically) to predict and propagate Gaussian mixtures [23]. Alternatively, non-parametric belief propagation methods like [26, 21] can be applied to solve *(c)*.

(2) We describe *conditional Bayesian mixture of experts* representations[2] that allow flexible discriminative modeling. These are based on hierarchical mixture of experts [14, 29, 6], an elaborated version of clusterwise or switching regression [9, 18], where the expert mixture proportions (called gates) are themselves observation-sensitive predictors, synchronized across experts to give properly normalized state distributions for any input observation. Inference is simple, contextual, and produces multimodal state conditionals. Our learning algorithm is different from the one of [29] in that we use sparse greedy approximations, and differs from [6] in that we use type-II maximum likelihood Bayesian approximations [15, 27], and not structured variational ones.

(3) We demonstrate the proposed algorithms on real and motion capture-based test sequences and present comparisons with nearest neighbor and regression methods.

2 Formulation

We work with discriminative graphical models with a chain structure, as shown in fig. 1, These have continuous temporal states $\mathbf{x}_t$, $t = 1 \dots T$, prior $p(\mathbf{x}_1)$, observations $\mathbf{r}_t$. For notational compactness, we also consider joint states $\mathbf{X}_t = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$ or joint observations $\mathbf{R}_t = (\mathbf{r}_1, \dots, \mathbf{r}_t)$. Learning and inference is based on local conditionals: $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, $p(\mathbf{x}_t|\mathbf{r}_t)$, and $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$.

2.1 Discriminative Density Propagation

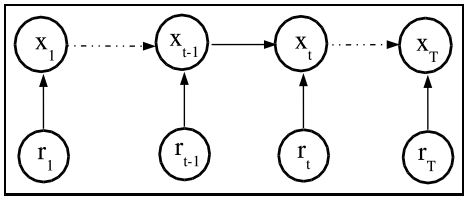

Figure 1: A discriminative chain model reverses the direction of the arrows that link the state and the observation, compared with a generative one. The state conditionals $p(\mathbf{x}_t|\mathbf{r}_t)$ or $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ can be learned using training pairs, and directly predicted during inference. Instead, a generative approach will model and learn $p(\mathbf{r}_t|\mathbf{x}_t)$ and do a more complex probabilistic inversion to compute $p(\mathbf{x}_t|\mathbf{r}_t)$ via Bayes' rule.

For filtering, we wish to compute the optimal distribution $p(\mathbf{x}_t|\mathbf{R}_t)$ for the state $\mathbf{x}_t$, conditioned by observations $\mathbf{R}_t$ up to time t. The filtered density can be derived as (see [24] for conditional independence assumptions implied by the graphical model in fig. 1 and for a proof):

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \qquad (1)$$

[2] An expert is any functional approximator, *e.g.* a perceptron or regressor.

(In fact, (1) can be derived more generally, based on a predictive conditional that depends on a larger window of observations up to time t [24].) In practice, we model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ as a conditional Bayesian mixture of M experts (*c.f.* §2.2). The prior $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is also represented as a mixture with M components. To compute the filtered posterior we integrate M^2 pairwise products of Gaussians analytically. This requires the linearization of our generally non-linear, but parametric, easily differentiable state conditionals. The means of the expanded posterior are clustered and the centers are used to initialize a reduced M-component approximation, refined using variational optimization [23].[3]

It is worth noticing that a discriminative corrective conditional $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ can be in practice more sensitive to incorrect previous state estimates than 'memoryless' distributions like $p(\mathbf{x}_t|\mathbf{r}_t)$. However we assume, as in any probabilistic approach, that the training and testing data are representative samples from the true underlying distributions in the domain. In practice, for improved robustness it is straightforward to include an importance sampler based on $p(\mathbf{x}_t|\mathbf{r}_t)$ to eq. (1), as we also use for initialization (see §3). Often it is also useful to correct out-of-sample observations $\mathbf{r}_t$ (caused *e.g.* by inaccurate silhouettes due to shadows) by projecting onto $p(\mathbf{r})$. Out of sample inputs or high entropy filtered posteriors can be indicative heuristics of the loss of track, or the absence of the target from the scene.

2.2 Bayesian Mixture of Experts Model (BME)

This section describes our methodology for learning multimodal conditional distributions for discriminative tracking (*e.g.* $p(\mathbf{x}_t|\mathbf{r}_t)$ or $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ in §2.1). Our proposal is motivated by the observation that many perception problems like reconstruction or tracking involve the recovery of inverse, intrinsically multivalued mappings. Static or dynamic state estimation ambiguities translate into multimodal conditional distributions (fig. 2). To represent them we use several 'experts' that are simple function approximators. The experts transform their inputs[4] into output predictions that are combined in a probabilistic mixture model based on Gaussians centered around them. The model is consistent across experts and inputs, *i.e.* the mixing proportions of the experts reflect the distribution of the outputs in the training set and they sum to 1 for every input. Some input domains can be predicted competitively by multiple experts and will have multimodal conditionals. Other 'unambiguous' inputs may be predicted by a single expert, with the others effectively switched-off, having negligible probability (see fig. 2). This is the rationale behind a Bayesian mixture of experts and provides a powerful mechanism for the contextual modeling of complex multimodal distributions. Formally this is described by:

$$p(\mathbf{x}|\mathbf{r}, \mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\lambda}) = \sum_{i=1}^{M} g(\mathbf{r}|\boldsymbol{\lambda}_i) p(\mathbf{x}|\mathbf{r}, \mathbf{W}_i, \boldsymbol{\Omega}_i^{-1}) \quad (3)$$

$$g(\mathbf{r}|\boldsymbol{\lambda}_i) = \frac{e^{\boldsymbol{\lambda}_i^\top \mathbf{r}}}{\sum_k e^{\boldsymbol{\lambda}_k^\top \mathbf{r}}} \quad (4)$$

$$p(\mathbf{x}|\mathbf{r}, \mathbf{W}_i, \boldsymbol{\Omega}_i^{-1}) = \mathcal{N}(\mathbf{x}|\mathbf{W}_i\mathbf{r}, \boldsymbol{\Omega}_i^{-1}) \quad (5)$$

Here $\mathbf{r}$ are input or predictor variables, $\mathbf{x}$ are outputs or responses, g are *input dependent* gates, computed in terms of regressors *c.f.* (4), with weights $\boldsymbol{\lambda}_i$. Notice how g are normalized to sum to 1 for consistency, by the softmax construction, for any given input $\mathbf{r}$. Also p are Gaussian distributions (5) with covariances $\boldsymbol{\Omega}_i^{-1}$, centered at different 'expert' predictions, here regressors[5] with weights $\mathbf{W}_i$. The parameters of the gates and the ones of the experts are collectively stored in $(\boldsymbol{\lambda}, \mathbf{W}, \boldsymbol{\Omega})$. As in many Bayesian settings [15, 27, 6], the weights $(\boldsymbol{\lambda}, \mathbf{W})$, are controlled by hierarchical priors, typically Gaussians with mean 0, and having inverse variance hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ controlled by a second level of Gamma distributions. This gives an automatic relevance determination mechanism [15, 27] that avoids overfitting, and encourages compact models with fewer non-zero weights for efficient prediction.

Our **learning** algorithm for the mixture of experts model is more complex, here we omit details due to space limitations (see [24]). As in many prediction problems we optimize the parameters $\boldsymbol{\theta} = (\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ to maximize the log-likelihood of a data set, $\mathcal{T} = \{(\mathbf{r}_i, \mathbf{x}_i) \mid i = 1 \ldots N\}$, *i.e.* the accuracy of predicting $\mathbf{x}$ given $\mathbf{r}$, averaged over the data distribution. For learning, a full Bayesian treatment would require integration over all parameters and hyperparameters. Because this is intractable, we design an iterative Bayesian EM algorithm based on type-II maximum likelihood [15, 27]. This uses Laplace approximation for the hyperparameters and analytical integration for the weights, which in this setting become Gaussian [15, 27].

Our algorithm proceeds as follows. In the E-step we estimate the posterior:

[3] It is possible to use a generative model, but express the propagation rules in terms of discriminative conditionals, in order to simplify inference [24]:

$$p(\mathbf{x}_t|\mathbf{R}_t) \propto \frac{p(\mathbf{x}_t|\mathbf{r}_t)}{p(\mathbf{x}_t)} \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \quad (2)$$

where $p(\mathbf{x}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1})$. Implementing (2) requires recursively propagating both $p(\mathbf{x}_t|\mathbf{R}_t)$ and $p(\mathbf{x}_t)$ (an equilibrium approximation could be precomputed), two mixture simplification levels, inside the integrand and outside it through the multiplication by $p(\mathbf{x}_t|\mathbf{r}_t)$ and a division by $p(\mathbf{x}_t)$ (see [24] for details).

[4] The 'inputs' can be either observations $\mathbf{r}_t$, when modeling $p(\mathbf{x}_t|\mathbf{r}_t)$ or observation-state pairs $(\mathbf{x}_{t-1}, \mathbf{r}_t)$ for $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$. The 'output' is the state throughout. Notice that temporal information will be considered when learning $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$.

[5] We write the mixture using linear models to avoid clutter, but it is straightforward to use non-linear kernels, as we do in many experiments [24].

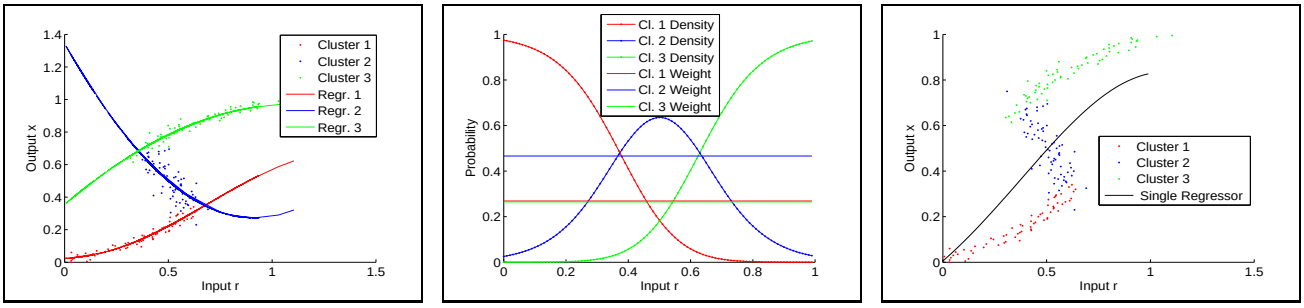

Figure 2: An illustrative dataset [6] consists of about 250 values of x generated uniformly in $(0, 1)$ and evaluated as $r = x + 0.3\sin(2\pi x) + \epsilon$, with ϵ drawn from a zero mean Gaussian with standard deviation 0.05. Notice that $p(x|r)$ is multimodal. *(a) Left* shows the data colored by the posterior membership probability h (6) of three expert kernel regressors. *(b) Middle* shows the gates g (4), as a function of the input, but also the three uniform probabilities (of the joint distribution) that are computed by a clusterwise regressor [9, 18]. *(c) Right* shows how a single kernel regressor cannot represent this multimodal distribution (it may either average the data or zig-zag through its multiple branches, depending on the kernel parameters).

$$h(\mathbf{x}, \mathbf{r}|\mathbf{W}_i, \mathbf{\Omega}_i, \boldsymbol{\lambda}_i) = \frac{g(\mathbf{r}|\boldsymbol{\lambda}_i)p(\mathbf{x}|\mathbf{r}, \mathbf{W}_i, \mathbf{\Omega}_i^{-1})}{\sum_j g(\mathbf{r}|\boldsymbol{\lambda}_j)p(\mathbf{x}|\mathbf{r}, \mathbf{W}_j, \mathbf{\Omega}_j^{-1})} \quad (6)$$

This gives the probability that the expert i has generated the data, and requires knowledge of both inputs and outputs (there is one h for each expert-training pair). In the M-step we solve two weighted regression problems, one for each expert and one for its gate. The first learns the expert parameters $\mathbf{W}_i$, based on training data $\mathcal{T}$, weighted according to the current membership estimates h. The second optimization teaches the gates g how to predict h.[6] Both solutions are based on ML-II, with greedy (regressor weight) subset selection. This strategy aggressively sparsifies the regressor by eliminating inputs with small weights after each iteration.

Inference (state prediction) is straightforward using (3). The result is a conditional mixture distribution with components and mixing probabilities that are input-dependent. In fig. 2 we explain the model using an illustrative toy example and show the relation with clusterwise and single regressors.

3 Experiments

This section describes our experiments as well as the training sets and the image features we use. We show results on real and artificially rendered motion capture-based test sequences, and give comparisons with existing methods.

Training Set, Model Representation and Image Features: It is difficult to obtain ground truth for human motion and even harder to train using many viewpoints or lighting conditions. Therefore, to gather data, we use as others [18, 20, 2, 28], packages like Maya (Alias Wavefront), with realistically rendered computer graphics human surface models, that we animate using human motion capture [1]. Our human representation ($\mathbf{x}$) is based on an articulated skeleton with spherical joints, and has 56 d.o.f. including global translation. Our database consists of about 3000 samples that involve a variety of human activities including walking, running, turns, gestures in conversations, quarreling and pantomime. We have studied empirically how ambiguous a sample of our training data is. This is shown and discussed in fig. 3.

Our choice of image features is based on previously developed methods for shape and texture modeling [8, 17, 5]. We work with silhouettes and we assume that in real settings these can be obtained using a statistical background subtraction method (we use one based on separately built foreground and background models, using non-parametric density estimation and motion segmentation). Silhouettes are informative for human pose estimation, although prone to certain ambiguities like the left / right limb assignment in side views or lack of observability of some of the d.o.f., *e.g.* 180^o ambiguities in the global azimuthal orientation for frontal views. These are multiplied by intrinsic forward / backward monocular ambiguities [25] that are common in many human interaction scenarios. (While no image descriptor set is likely to easily help discriminate them, this further motivates our probabilistic, multiple hypothesis approach.) We use shape context features extracted on the silhouette [5, 17, 2] (5 radial bins, 12 angular bins, with bin size range 1 / 8 to 3 on log scale). We also experiment with pairwise edge angle and distance histograms [3] collected inside the silhouette. The features are computed at a variety of scales and sizes for points sampled on the silhouette. To work in a common coordinate system, we cluster all features in the training set into K=40 clusters. To compute the representation of a new shape feature (a point on the silhouette), we 'project' onto the common basis by inverse distance weighted voting into the cluster centers. To obtain the representation ($\mathbf{r}$) for a new silhouette we regularly sample about 100 points on it's contour and add all their

[6] Prediction based on the input *only* is essential for inference, where membership probabilities (6) cannot be computed because the output is missing.

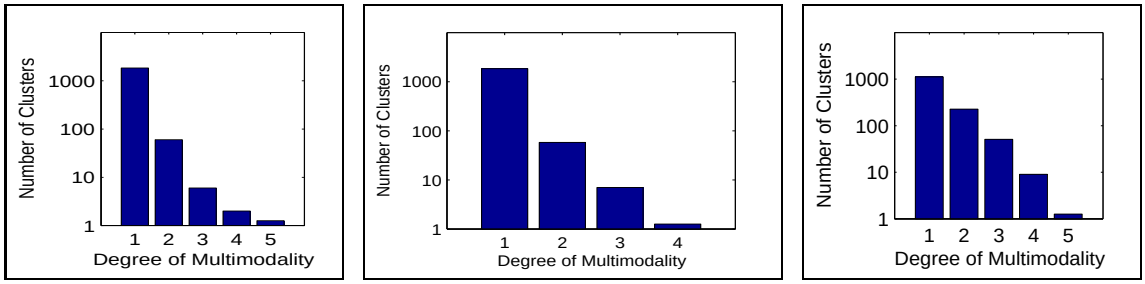

Figure 3: Analysis of 'multimodality' for a training set (the 'number of clusters' axis on logscale): *(a) Left:* $p(\mathbf{x}_t|\mathbf{r}_t)$ (1912 clusters / 2000 points). *(b) Middle:* $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ (1912 clusters / 2000 points). *(c) Right:* $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ (1409 clusters / 2000 points). We cluster the features and joint angle vectors, independently, into a large number of clusters. We build histograms for the number of joint angle clusters that fall under the same feature cluster. This quantifies the ambiguity in the database at the feature and joint angle cluster scale. We select many clusters to simulate the effect of small perturbations in the input. In those cases any feature neighbor (not necessarily the desired one) may be the closest to an input silhouette query. The input neighborhood induces a distribution over clusters of joint angles. We notice that even at this fine scale, the conditionals are multimodal. Decreasing the number of clusters in *(c)* sharply increases multimodality. Working with the previous state and the current observation (middle and right plot) does not eliminate ambiguity. This is not wild, but severe enough to cause tracking failure or significant errors during initialization (so we observe in various tests). We expect increasing ambiguity for larger training sets. A similar two-level clustering strategy is used to initialize the learning procedure for the BME models. We initially cluster based on the input components and then separately cluster the samples within each 'input' cluster based on the output components. This aims to avoid situations where single experts would inconsistently represent multiple branches of the inverse pose mapping (see fig. 2) leading to models that correspond to poor likelihood optima.

feature vectors into a feature histogram. This representation is semi-local, rich and has been effectively demonstrated in many applications, including texture recognition [8] or pose prediction [17, 20, 2].

Comparisons: We compare our Bayesian mixture of experts (BME) conditional models with other competing methods like weighted nearest neighbor (NN) or the relevance vector machine (RVM) [27]. Our test set consists of a variety of human activities obtained using motion-capture and artificially rendered. This provides ground truth and allows us to concentrate on the algorithms and factor out the variability given by the imperfections of our human model, or the noise in the silhouette extraction in real images. The results are shown and discussed in (the caption of) table 1. In general the BME model gives better average estimates and significantly lower maximum errors. Notice also that for the purpose of the comparison we have only considered the most probable prediction of the BME. However, while the correct solution is not always predicted as the most probable, it is often the case that is still present among the top modes predicted by the BME, see *e.g.* fig. 4c. For probabilistic tracking, this 'approximately correct' behavior is desirable because the correct solution will still be propagated with significant probability. We notice that despite transient distraction, the model often recovers the most probable solution during subsequent frames.

Real Image Sequences. Picking and Dancing: In fig. 5, we show the result of tracking a real image sequence consisting of 2 seconds of video, 60 fps. Our experiments involve both Bayesian single hypothesis tracking based on a single expert, propagated using (1), as well as multiple hypotheses tracking based on a BME model learned using 5 experts that are regressors with RBF kernels and degree of sparsity varying between 5%-25%. We initially tested the single hypothesis tracker. This tracks the beginning of the sequence but fails shortly after, as its input kernels stop firing due to an out-of-range input predicted from the previous timestep (see [24] for images and details). To factor out the effect of imperfect silhouettes or initialization[7] and to make sure that failure is due to motion or feature representation ambiguities, we also attempted to reconstruct a similar sequence using artificially rendered images, generated from a motion trajectory in the database. Even in that case, the single hypothesis tracker failed. In fig. 5 we show results using a 5-mode BME tracker that successfully reconstructs the motion. While the reconstruction is perceptually plausible, there are imperfections, possibly reflecting the bias introduced by our training set – *e.g.* notice that the knee of the model is tilted outward whereas the knee of the human is tilted inward. We also observe persistent multimodality for those joints more actively moving, *e.g.* the right wrist, the right femur and the right shoulder, which have, quite constantly, about 5 modes in their posterior. In general, in the beginning of the sequence there is more ambiguity for almost all the joints, but it tends to decrease (but not disappear) during tracking. However, the joints that are occluded or very much project inside the silhouette tend to have persistent ambiguities perhaps due to folds in clothing, or shadows.

[7] We initialize using the conditional $p(\mathbf{x}_t|\mathbf{r}_t)$, learned using BME. For single hypothesis tracking, we select the most probable component.

Sequence	$p(\mathbf{x}_t \mid \mathbf{r}_t)$			$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{r}_t)$		
	NN	RVM	BME	NN	RVM	BME
NORMAL WALK	4 / 20	2.7 / 12	2 / 10	7 / 25	3.7 / 11.2	2.8 / 8.1
COMPLEX WALK	11.3 / 88	9.5 / 60	4.5 / 20	7.5 / 78	5.67 / 20	2.77 / 9
RUNNING	7 / 91	6.5 / 86	5 / 94	5.5 / 91	5.1 / 108	4.5 / 76
CONVERSATION	7.3 / 26	5.5 / 21	4.15 / 9.5	8.14 / 29	4.07 / 16	3 / 9
PANTOMIME	7 / 36	7.5 / 53	6.5 / 25	7.5 / 49	7.5 / 43	7 / 41

Table 1: Comparative results showing RMS errors per joint angle (average error / maximum joint average error) in degrees for two conditional models, $p(\mathbf{x}_t|\mathbf{r}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$. We compare three different algorithms on motion-capture, synthetically generated test data (we select the best candidate for each test input, there is no probabilistic tracking, but $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ has memory). The algorithms are: NN (nearest neighbor with soft state weighing, proportional to the inverse distance to input feature), RVM (relevance vector machine), BME (Bayesian mixture of experts, with most probable mode selected). We use several training sets: walking diagonal w.r.t. to the image plane (train 300, test 56), complex walking towards the camera and turning back (train 900, test 90), running parallel to the image plane (train 150, test 150), conversation involving some hand movement and turning (train 800, test 160), pantomime (1000 train, 100 test). The training has been done separately for each sequence, to limit ambiguity, and we initialize from ground truth. This favors unimodal approaches, especially when using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$, as they may not recover from an incorrect initialization. Notice that BME has typically smaller average errors and significant smaller maximum errors. The large maximum error for running seems consistent across various methods and corresponds to the right hand joint. For the BME, we only measure the error w.r.t. the most probable mode but we observe that even when this is not the most likely, it is still among the probable ones predicted (see fig. 4c).

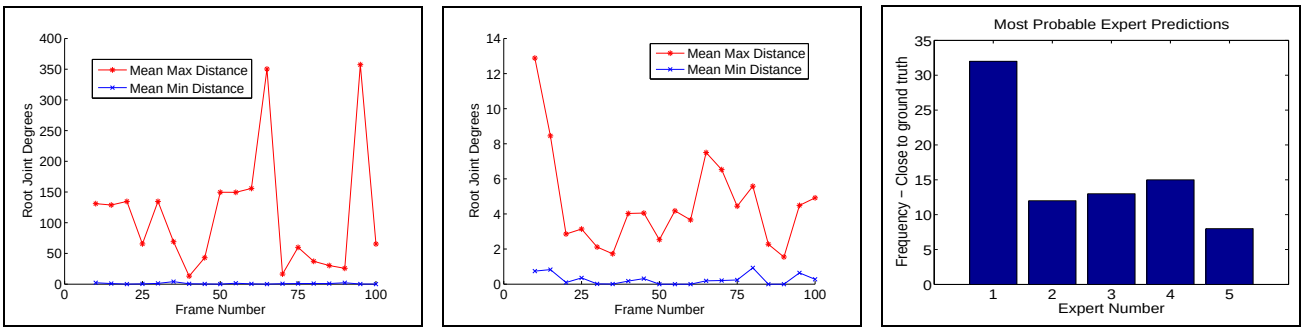

Figure 4: Quantitative tracking results for the dancing sequence. *(a) Left:* shows the maximum and minimum distance for the modes of the root joint vertical axis rotation angle. The minimum distance is only informatively shown, it does not necessarily reflect modes that will survive the mixture simplification. Most likely, modes that cluster together will collapse. *(b) Middle:* same as *(a)* for the left femur. *(c) Right:* shows the accuracy of our mixture predictor. Notice that the most probable mode according to the model is not always the most accurate one.

We conclude with experiments where we reconstruct a more challenging dancing sequence consisting of 400 frames. We include 300 frames in the training set and test on 100 of them. Image results are shown in fig. 6, whereas quantitative results are shown in fig. 4. Although the poses we reconstruct are not geometrically perfect and there are sometimes errors at the hands and the legs, the BME prediction still captures the underlying 3d poses in a perceptually plausible way.

4 Conclusions

We have presented a mixture density propagation framework for temporal inference using discriminative models. We argued that despite their success, existing methods do not offer a formal management of uncertainty and we explained why current representations cannot model multivalued relationships that are pervasive in inverse, perception problems. We contribute by establishing the density propagation rules in discriminative, continuous, temporal chain models and by proposing compact Bayesian mixture of experts models capable of representing multimodal conditionals. We show results on real and synthetically generated image sequences, and give comparisons against nearest neighbor and regression methods. Our study suggests that flexible conditional modeling and uncertainty propagation are both essential for successful reconstruction. We hope that this research will bring discriminative and generative tracking algorithms closer and help

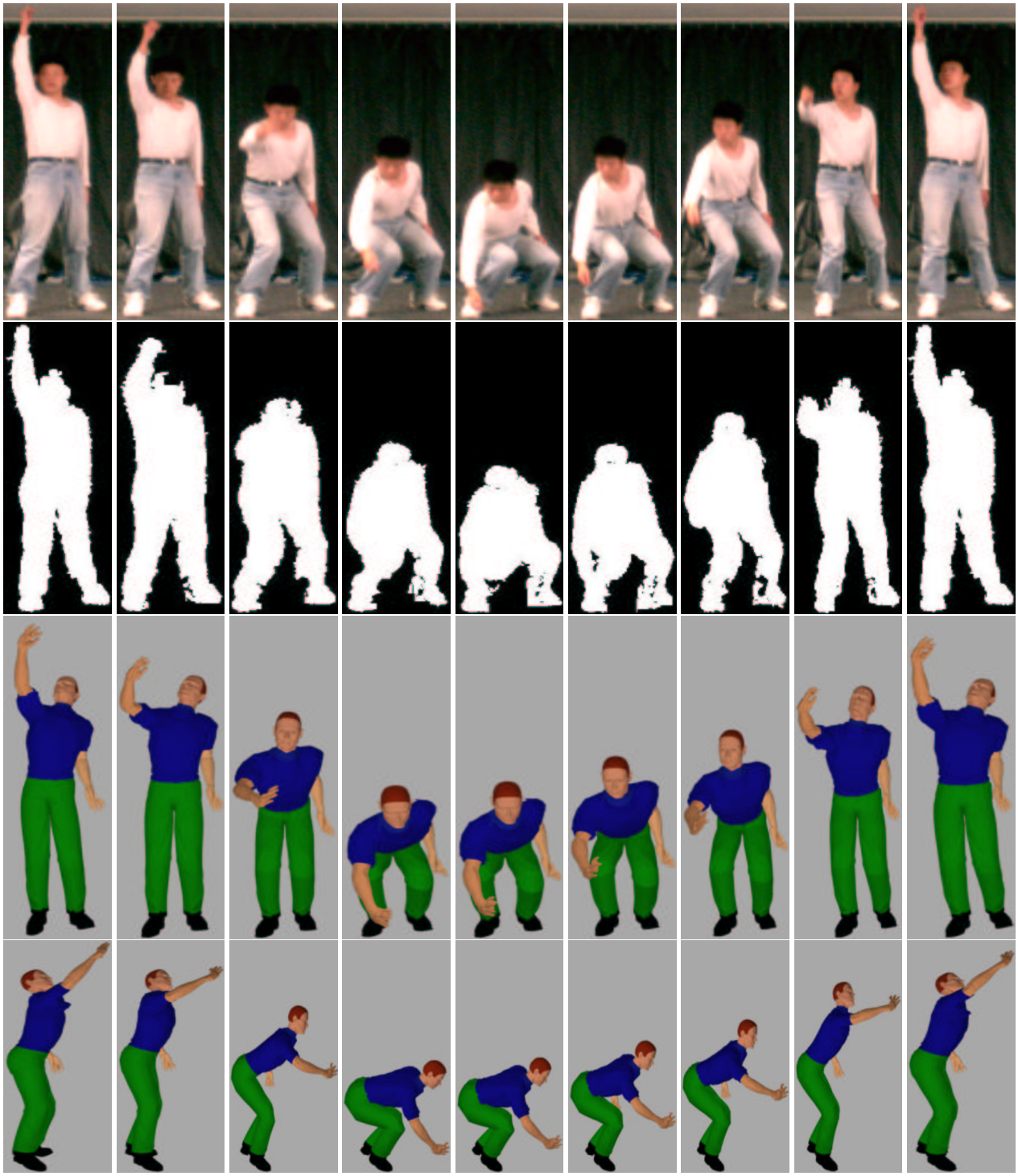

Figure 5: *First row:* Original image sequence. *Second row:* Image silhouettes. *Third row:* Reconstruction seen from the same viewpoint used for training, *Fourth row:* Reconstruction seen from a synthetic viewpoint. Notice that despite noisy silhouettes, our probabilistic tracker based on Bayesian mixture of experts (BME) conditionals can reconstruct the motion with reasonable perceptual accuracy (however, there are imperfections, *e.g*. the right knee of the subject is tilted inward, whereas the one of the model is tilted outward). A single hypothesis Bayesian tracker fails on the same sequence (see [24]).

Figure 6: Tracking and 3d reconstruction of a dancing sequence. *(a) Top* row shows original images and silhouettes; *(b) Bottom* row shows reconstructions from training (left) and new synthetic viewpoint (right).

stimulate a fruitful debate on their relative advantages within a common probabilistic framework.

Future Work: We plan to do a detailed sensitivity analysis w.r.t. motions and shapes that deviate from the training set. We also study alternative, more compact state and feature representations based on dimensionality reduction, and investigate scaling aspects for large motion capture databases. Other research directions involve reconstructing multiple people and handling occlusions and different observation representations.

References

[1] CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003.

[2] A. Agarwal and B. Triggs. 3d human pose from silhouettes by Relevance Vector Regression. In *CVPR*, 2004.

[3] F. Aherne, N. Thacker, and P. Rocket. Optimal pairwise geometric histograms. In *British Machine Vision Conference*, 1997.

[4] A. Athistos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *ICCV*, 2003.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24, 2002.

[6] C. Bishop and M. Svensen. Bayesian mixtures of experts. In *UAI*, 2003.

[7] M. Brand. Shadow Puppetry. In *ICCV*, pages 1237–44, 1999.

[8] O. Cula and K. Dana. 3D texture recognition using bidirectional feature histograms. *IJCV*, 59(1):33–60, 2004.

[9] W. DeSarbo and W. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, (5):249–282, 1988.

[10] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *CVPR*, 2000.

[11] K. Grauman, G. Shakhnarovich, and T. Darell. Inferring 3D structure with a statistical image-based shape model. In *ICCV*, 2003.

[12] N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. *NIPS*, 1999.

[13] M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *IJCV*, 1998.

[14] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, (6):181–214, 1994.

[15] D. Mackay. Bayesian interpolation. *Neural Computation*, 4(5):720–736, 1992.

[16] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *ICML*, 2000.

[17] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, 2002.

[18] R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *NIPS*, 2002.

[19] S. Roth, L. Sigal, and M. Black. Gibbs Likelihoods for Bayesian Tracking. In *CVPR*, 2004.

[20] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *ICCV*, 2003.

[21] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking Loose-limbed People. In *CVPR*, 2004.

[22] C. Sminchisescu and A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *ICML*, pages 759–766, Banff, 2004.

[23] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *CVPR*, volume 2, pages 608–615, Washington D.C., 2004.

[24] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Learning to reconstruct 3D human motion from Bayesian mixtures of experts. A probabilistic discriminative approach. Technical Report CSRG-502, University of Toronto, October 2004.

[25] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *CVPR*, volume 1, pages 69–76, Madison, 2003.

[26] E. Sudderth, A. Ihler, W. Freeman, and A.Wilsky. Nonparametric belief propagation. In *CVPR*, 2003.

[27] M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *JMLR*, 2001.

[28] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *ICCV*, 2003.

[29] S. Waterhouse, D.Mackay, and T.Robinson. Bayesian methods for mixtures of experts. In *NIPS*, 1996.