Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference

Cristian Sminchisescu
TTI-C
crismin@nagoya.uchicago.edu
http://ttic.uchicago.edu/~crismin

Atul Kanaujia and Dimitris Metaxas
Rutgers University
{kanaujia,dnm}@cs.rutgers.edu
http://www.cs.rutgers.edu/~kanaujia,~dnm

Abstract

We present an algorithm for jointly learning a consistent bidirectional generative-recognition model that combines top-down and bottom-up processing for monocular 3d human motion reconstruction. Learning progresses in alternative stages of self-training that optimize the probability of the image evidence: the recognition model is tunned using samples from the generative model and the generative model is optimized to produce inferences close to the ones predicted by the current recognition model. At equilibrium, the two models are consistent. During on-line inference, we scan the image at multiple locations and predict 3d human poses using the recognition model. But this implicitly includes one-shot generative consistency feedback. The framework provides a uniform treatment of human detection, 3d initialization and 3d recovery from transient failure. Our experimental results show that this procedure is promising for the automatic *reconstruction of human motion in more natural scene settings with background clutter and occlusion.*

1. Introduction

Analyzing three-dimensional human motion in real world environments is an actively growing field with a broad scope of applications spanning video browsing and indexing, human-computer interaction and surveillance. The problem has been traditionally attacked using the powerful machinery of top-down, *generative modeling*, with image-based feedback provided within an analysis-by-synthesis loop. Despite being a natural way to model the appearance of complex articulated structures, the success of generative models has been partly shadowed because it is computational demanding to infer the distribution on their hidden states (here human joint angles) and because their parameters are unknown and variable across many real scenes.

The difficulty of generative modeling has motivated the advent of a complementary class of bottom-up, feed-forward, discriminative *recognition methods* which predict state distributions directly from image features. Despite being simple to understand and fast, recognition methods tend to assume that the object of interest is segmented and could be blind-sighted by the lack of feedback – they cannot self-asses accuracy. It may be possible to bootstrap them using powerful global or part-based 2d human detectors, but one challenge is to make this approach scale to the range of poses needed for general 3d reconstruction without progressively modeling most of the 3d constraints. This is partly because the space of (even typical) human articulation is large, because self-occlusion and foreshortening are hard to model in 2d, and because most of the human body parts do not have the distinctiveness that allows their unambiguous detection and combination.

To summarize, what appears to be necessary is a mechanism to consistently integrate top-down and bottom-up processing: the flexibility of 3d generative modeling (represent a large set of possible poses of human body parts, their correct occlusion and foreshortening relationships and their consistency with the image evidence) with the speed and simplicity of feed-forward processing.

In this paper we present one possible way to meet these requirements based on a bidirectional model with both recognition and generative sub-components. *Learning* the model parameters alternates self-training stages in order to maximize the probability of the observed evidence (images of humans). During one step, the recognition model is trained to invert the generative model using samples drawn from it. In the next step, the generative model is trained to have a state distribution close to the one predicted by the recognition model. At local equilibrium, which is guaranteed, the two models have consistent, registered parameterizations. During *on-line inference*, the estimates are driven mostly by the fast recognition model, but implicitly include one-shot generative consistency feedback.

The resulting 3d temporal predictor operates similarly to existing 2d object detectors. It searches the image at different locations and uses the recognition model to hypothesize 3d configurations. Feedback from the generative model

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

helps to downgrade incorrect competing 3d hypotheses and to decide on the detection status (human or not) at the analyzed image sub-window. Our results obtained in *monocular* video show that the proposed model is promising for the *automatic* reconstruction of 3d human motion in environments with background clutter. The framework provides a uniform treatment of human detection, 3d initialization and 3d recovery from transient failure.

1.1. Related Work

The research we present relates to the growing body of work in the area of 2d human detection, 3d motion reconstruction, as well as learning and approximation. Due to space limitations, we only give a brief survey without aiming at a full literature review. Generative human models and efficient inference algorithms are described in [6, 24, 25, 26, 30]. Discriminative 3d reconstruction methods, based on single or multiple hypotheses, in both static and temporal settings are proposed in [22, 23, 2, 7, 28]. Systems based on global [17, 4, 31, 17] or part-based human detectors [19, 21, 15] have been demonstrated convincingly, and can be an alternative 2d front end to the integrated 3d reconstruction scheme proposed here.

Generative models and variational learning algorithms for extracting motion layers in video have been explored by [11] and recently by [14] with excellent results. Generative models based on variational techniques [8] have been investigated by [20] in the context of switching linear dynamical models for human motion analysis. Learning flexible aspect graphs of 2d human exemplars is a goal in the work of [29] where specific models and inference algorithms are given.

The combination of generative and discriminative 3d human models for the static case has been studied by [22] who has employed a mixture of neural networks for the recognition model and a neural network model for verification. This type of approach has been extended to the dynamic case [5] in a multi-camera setting, in conjunction with Parameter Sensitive Hashing, a fast nearest neighbor method [23] used to initialize model-based non-linear optimization. The combination of top-down and bottom-up information is considered in the work on generative models estimated using importance sampling [10, 6], although their proposal distributions are typically predefined and not learned, being good starting points for inference rather than completely plausible solutions to it. Combining feed-forward and feedback information is also a research focus in the biologically inspired machine learning community. The wake-sleep algorithm [9] can be viewed as a notable precursor of the one we propose, although it is based on a significantly different model architecture, cost function and optimization procedure for learning.

2. Modeling and Learning

In the section we describe the generative and the recognition models that we use and propose a self-supervised algorithm that can jointly and consistently learn them.

2.1. Generative Model

Consider a non-linear generative model $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})$ with hidden state $\mathbf{x}$ with $d = \dim(\mathbf{x})$, observation $\mathbf{r}$, and parameters $\boldsymbol{\theta}$. To simplify notation, but without loss of generality, we assume a uniform (normalized) state prior over the domain and a robust observation model:

$$p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}) = (1-w)\cdot\mathcal{N}(\mathbf{r};\mathcal{G}(\mathbf{x}),\Sigma_{\boldsymbol{\theta}}) + o_{\boldsymbol{\theta}}\cdot w \tag{1}$$

This corresponds to a mixture of a Gaussian having mean $\mathcal{G}(\mathbf{x})$ and covariance $\Sigma_{\boldsymbol{\theta}}$, and a uniform background of outliers $o_{\boldsymbol{\theta}}$ with proportions given by w. The outlier process is truncated at large values, so the mixture is normalizable.

In our case, the state space $\mathbf{x}$ represents human joint angles, the parameters $\boldsymbol{\theta}$ include the Gaussian observation noise covariance and the weighting of outliers (the human body proportions are fixed in the experiments, but they could be also learned, in principle). $\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x})$ is a complex non-linear transformation that predicts human contours and internal edges (it includes non-linear kinematics, occlusion analysis and perspective projection). The edge image is based on SIFT [18] descriptors $\mathbf{r}$, densely computed at a regular grid inside the target detection window and concatenated in a descriptor vector.

For simplicity, we will also use an equivalent, energy-based model representation:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}) &= \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{x})}\exp(-E_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x})) && (2)\\ E_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}) &= -\log[(1-w)\mathcal{N}(\mathbf{r};\mathcal{G}(\mathbf{x}),\Sigma_{\boldsymbol{\theta}}) + o_{\boldsymbol{\theta}}w] - && (3)\\ &\qquad - \log Z_{\boldsymbol{\theta}}(\mathbf{x}) && (4) \end{aligned}$$

with normalization constant $Z_{\boldsymbol{\theta}}(\mathbf{x}) = \int_{\mathbf{r}}\exp(-E_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))$ that can be easily computed by sampling from the mixture of Gaussian and uniform outlier distribution. Using Bayes rule, the model state conditional is: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{r})}\exp(-E_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))$, but computing $Z_{\boldsymbol{\theta}}(\mathbf{r}) = \int_{\mathbf{x}}\exp(-E_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))$ is intractable because the average is taken w.r.t. the unknown state distribution.

2.2. Recognition Model

For state inference tasks, one can in principle work only with a generative model, but this is computationally expensive. The common experience [9, 22, 2, 7, 19, 28] (besides biological arguments) seems to indicate that for speed and robustness it is *also* useful to consider a diagnostic (discriminative) model, designed to condition on the observation, not to generate it. Strictly, this is a generative model of the state, a fast cache that inverts the complex observation model and operates in conjunction with it. The recognition model we use is a *conditional* mixture of Bayesian experts, sparse function approximators which know how to invert certain ranges *but not the entire domain* of their input (observation). While the observation-to-state mapping is ambiguous (multivalued) in general, the experts cooperate to compute an approximation to $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r})$ that is valid globally. In this sense, the conditional model is a machinery for representing multimodal distributions contextually:

$$Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) = \sum_{i=1}^{M} g_i(\mathbf{r})\mathcal{N}(\mathbf{x}; F_i(\mathbf{r}) = \mathbf{W}_i\mathbf{r}, \boldsymbol{\Omega}_i) \quad (5)$$

$$g_i(\mathbf{r}) = \frac{\exp(\boldsymbol{\lambda}_i^\top \mathbf{r})}{\sum_{k=1}^{M} \exp(\boldsymbol{\lambda}_k^\top \mathbf{r})} \quad (6)$$

where $g_i(\mathbf{r})$ are observation dependent *feedforward gates*, computed using linear regressors with weights $\boldsymbol{\lambda}_i$. The gates are normalized in order to consistently sum to 1, using the softmax function, for any input $\mathbf{r}$. The experts are Gaussian distributions (5) with diagonal covariances $\boldsymbol{\Omega}_i$, centered at different expert predictions $F_i(\mathbf{r})$. The approximators are sparse Bayesian regressors with weights $\mathbf{W}_i$. The joint parameter vector is $\boldsymbol{\nu} = \{(\boldsymbol{\lambda}_i, \mathbf{W}_i, \boldsymbol{\Omega}_i) \,|\, i = 1 \ldots M\}$.

A recognition model given by (5) has been successfully used for high-dimensional human pose prediction problems based on silhouettes [28]. Nevertheless, the inherent feedforward processing is prone to 'hallucination'. Due to the lack of feedback, the model can produce either incorrect states or incorrect probability estimates or both without ever knowing it. This raises the question whether additional problem structure can be considered for consistency feedback.

In this paper we show a simple but effective gate construction (to our knowledge not proposed in the hierarchical mixture of experts literature [13, 3]) based on an approximate generative observation model. Such models exists for many visual processes, being often used to generate (quasi-synthetic) data for training feedforward models [22, 28]. The information indirectly captured by the gates can be further refined by the generative model which allows to verify that a 3d state, inferred bottom-up based on an observation, can in turn, stochastically generate a prediction close to to it. However, when multiple predictions are equally likely, the relative expert probabilities are driven by the data density, as before (6). This leads to a natural product *feedback gate* function where the generative observation model reweights the *feedforward gate*:

$$g_i(\mathbf{r}) = \frac{\exp[-E_{\boldsymbol{\theta}}(\mathbf{r}|F_i(\mathbf{r})) + \boldsymbol{\lambda}_i^\top \mathbf{r}]}{\sum_{k=1}^{M} \exp[-E_{\boldsymbol{\theta}}(\mathbf{r}|F_k(\mathbf{r})) + \boldsymbol{\lambda}_k^\top \mathbf{r}]} \quad (7)$$

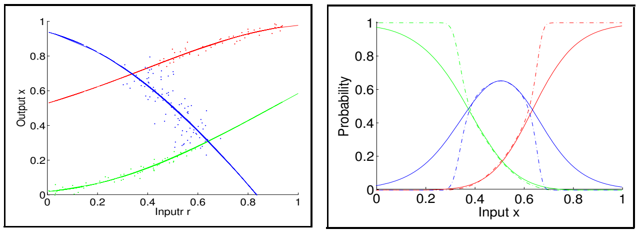

Figure 1. **Best viewed in color and zoomed**. Reworking a dataset [3, 28] which consists of 250 values of x generated uniformly in $(0, 1)$ and evaluated as $r = x + 0.3\sin(2\pi x) + \epsilon$, with ϵ drawn from a zero mean Gaussian with standard deviation 0.05. *(a) Left* shows the data colored by the posterior membership probability of three expert kernel regressors (also shown). *(b) Right* shows the *feed-forward gates* of [28] (linear regressors) as a function of the input, but also the *feedback gates* (**dashed**) computed using the generative model, *c.f.* (7) (the experts in *(a)* are identical in both cases). The feed-forward gates are well fitted (also consistent with the results of [3, 28]), but notice the *large* difference in the probability estimate between those and the generative gates in the range (0,.35) and (.65,1). This shows the exceptional impact of feedback provided by the generative observation model. The data sampled from a model with feed-forward gates may not always be accurately distributed, especially in transition regions where the experts change relative strength. The generative gates make the conditional model sharper. See, *e.g.*, predictions for $r \approx 0.3$ when the lowest (green) expert should dominate. However, in the feedforward model, about 0.2 of the probability value leaks to the middle expert (blue).

By construction, the conditional model in (5), (7) integrates bottom-up and top-down information consistently. See fig. 1 and its related discussion for an illustration.

2.3. Learning a Generative-Recognition Tandem

An important research problem for visual processing is to jointly learn a consistent generative-recognition model pair, that is useful for inference. We propose an algorithm for model learning (including choices of tractable approximating distributions), based on Variational Expectation-Maximization (VEM) [12].

The goal of both learning and inference is to maximize the probability of the evidence (observation) under the data generation model:

$$\log p_{\boldsymbol{\theta}}(\mathbf{r}) = \log \int_{\mathbf{x}} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r}) = \log \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})}{Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})} \quad (8)$$

$$\geq \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})}{Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})} = KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})) \quad (9)$$

which is based on Jensen's inequality [12], and KL is the Kullback-Leibler divergence between two distributions. For learning, (8) will sum over the observations in the training set, omitted here for clarity. *The recognition model $Q_{\boldsymbol{\nu}}$ acts as an approximating variational distribution for the generative model.* This is the same as maximizing a lower bound

Algorithm for Bidirectional Model Learning

E-step: $\boldsymbol{\nu}^{k+1} = \arg\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\theta}^k)$
Train the *recognition* model using samples from the current *generative* model.

M-step: $\boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\nu}^{k+1}, \boldsymbol{\theta})$
Train the *generative* model to have state posterior close to the one predicted by the current *recognition* model.

Figure 2. Variational Expectation-Maximization (VEM) algorithm for jointly learning a generative and a recognition model.

on the log-marginal (observation) probability of the generative model, with equality when $Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r})$.

$$\log p_{\boldsymbol{\theta}}(\mathbf{r}) - KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r})) \quad (10)$$

$$= KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})) \quad (11)$$

According to (8) and (10), optimizing a variational bound on the observed data is equivalent to minimizing the KL divergence between the state distribution inferred by the generative model $p(\mathbf{x}|\mathbf{r})$ and the one predicted by the recognition model $Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})$. This is equivalent to minimizing the KL divergence between the recognition distribution and the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})$ – the cost function we work with:

$$KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})) = -\int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \quad (12)$$

$$+ \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) = \mathcal{L}(\boldsymbol{\nu},\boldsymbol{\theta}) \quad (13)$$

Notice that the cost $\mathcal{L}(\boldsymbol{\nu},\boldsymbol{\theta})$ balances two conflicting goals: assign values to states that have high probability under the generative model (the second term), but at the same time be as uncommitted as possible (the first term measuring the entropy of the recognition distribution). The gradient-based learning algorithm we use is summarized in fig. 2 and is guaranteed to converge to a locally optimal solution for the parameters.

The procedure is, in principle, self-supervised (one has to only provide the image of a human *without* the corresponding 3d human joint angle values), but we initialize by training the recognition and the generative models separately. We use supervised methods based on motion capture data (3d joint angle states $\mathbf{x}$) paired with noisy SIFT [18] image descriptors (observations $\mathbf{r}$).[1] These are gathered inside bounding boxes of artificially rendered human silhouettes, placed on backgrounds drawn from a set of natural images (see §3).

Inference using a mixture of mean field distributions: For efficient learning, it is critical that the variational family of distributions $Q_{\boldsymbol{\nu}}$ has a tractable form, because their expectations are computed repeatedly. We design the recognition model $Q_{\boldsymbol{\nu}}$ (§2.2) as a conditional Gaussian mixture, with means and diagonal covariances learned by training independent kernel regressors for each state dimension. Each mixture component thus factorizes as a product of univariate Gaussians: $\mathcal{N}(\mathbf{x}; F_i(\mathbf{r}), \mathbf{\Omega}_i) = \prod_{j=1}^{d} \mathcal{N}(x_j; F_i^j(\mathbf{r}), \omega_{ij})$, where ω_{ij} is the expert i's variance for state dimension j. This considerably simplifies the calculation of expectations needed for learning. The resulting variational approximation is a mixture of mean-field distributions. See [12] for a detailed derivation of the iterative parameter updates for this class of approximations.

The approximation accuracy may be of legitimate concern given the strong correlations among the human state variables. This turns out to be effective because, even though different dimensions of the state space are decoupled, they are strongly constrained by the observation $\mathbf{r}$ (through $F_i(\mathbf{r})$), by the Bayesian expert parameters, and indirectly by feedback from the generative model.[2] An alternative for strongly coupled models is to first decorrelate the state dependency using a non-linear dimensionality reduction method [27], and then use weaker variational approximations on the latent representation (*e.g.* a mean field procedure based on factorized state distributions *not* constrained by the observation [12]). This is computationally more attractive, but we found it was less accurate ($\approx 3^o$ higher average error per human joint angle in preliminary tests). It also comes at the price of operating with models having a restricted capacity. Here we aim at a balance between specificity and speed (the recognition model) and representation generality (the generative model).

Online inference (3d reconstruction and tracking) is straightforward using the E-step in fig. 2, but for efficiency we usually work only with the recognition model ((5) and (7)). More generally, it is possible to iterate the mean field updates (do generative inference) when the recognition distribution has high entropy. The model then effectively switches between a discriminative density propagation rule [28] but with *generative gates c.f.* (5) and (7), and a generative propagation rule [10, 6, 24, 26]. This offers a natural 'exploitation-exploration' or prediction-search tradeoff.

3. Experiments

We describe 3d reconstruction experiments in both real image sequences and in quasi-synthetic ones (combination of real and synthetic images, with known 3d ground truth).

[1] A worth-studying alternative with good invariance properties are the powerful convolutional networks of [16].

[2] This factorial state construction is not exclusively driven by tractability constraints. Models based on experts learned separately for each state dimension have been shown to give jointly consistent predictions [2, 28].

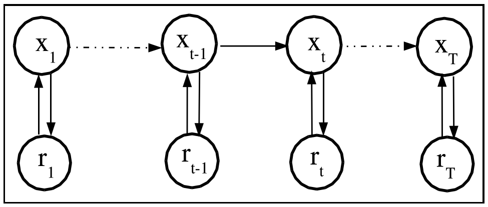

Figure 3. A bidirectional model for learning and recognition. During learning, the model is symmetrically driven by both recognition (bottom-up) and generative (top-down) connections. It alternates by learning one directional component at a time. During online inference, the model is driven mostly by recognition connections, but includes one-shot (generative) consistency feedback.

State representation and image descriptors: The state variables ($\mathbf{x}$) are 56d human joint angles. The image features ($\mathbf{r}$) are vectors of SIFT descriptors [18], computed at a densely sampled grid of image locations [31, 4], inside a putative detection window. (For training, this is a scaled bounding box of the human silhouette.) Each descriptor consists of 9 blocks (3x3), 6x6 pixels, with 8 bin gradient histograms, obtained by orientation voting using bilinear interpolation. The descriptor blocks have no overlap. The detection window size is 252x144 that results in an observation descriptor vector of size 8064. The choice of this image descriptor (here referred as BSIFT) is partly motivated by the need for robustness to clutter. Another choice used successfully in the past is a histogram descriptor (of SIFT or shape contexts) [28, 2] (here referred as HSIFT). The SIFT responses are clustered to obtain a codebook and each new detection window is represented as a histogram w.r.t. it (*e.g.* by vector quantizing using soft voting into the nearby bins). However, we found that this encoding tends to spread the background distribution among all the components of the descriptor, making noise suppression (*e.g.* by sparsification) difficult, see fig. 4. Using this descriptor the recognition model couldn't generalize, see table 1.

Database and learning: The learning algorithm we propose is, in principle, *self-supervised*, but given the large number of parameters (the distribution of observations for the generative model, the expert parameters for the recognition model), a consistent initialization is critical for good learning performance. Hence, we start by training the generative and the recognition model separately using *supervised* procedures. We use a computer graphics human body model (from [28]), animated based on a database of human motion capture [1]. We use 5500 samples that include activities like walking, running and some human conversations (4500 for training and 1000 for testing). Each pose in the database is rendered from several viewpoints. This training strategy has been effective for 3d pose prediction based on silhouettes [22, 23, 28, 2], but we find it cannot directly generalize to more complex scenes where distracting background clutter and occlusion are prevalent. (The interested reader is advised to also review the experiments of [2, 23] for alternative edge-based observation encodings applied to 3d human pose estimation in cluttered images.)

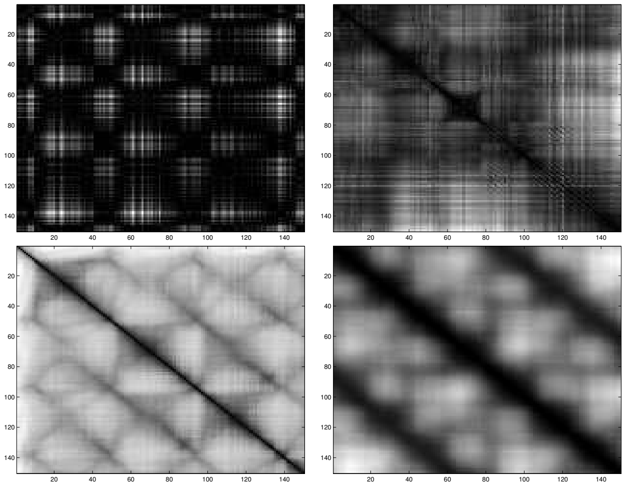

Figure 4. Affinity matrices for different feature types for walking seen sideways, under different backgrounds: top row shows a 128d HSIFT representation, without and with background clutter. Notice that the background distribution is spread in the entire descriptor. Bottom row shows a 8064d BSIFT descriptor, but where the components with zero weights after training have been eliminated (see text). The descriptor is better, but useful correlations in the sub-diagonal blocks are still suppressed. Bottom-right shows the affinity matrix of the 3d joint angles.

To promote diversity, we enhance the training set by placing the synthetic sprites on realistic backgrounds drawn from a database of 50 natural images (see fig. 5). In order to model variability due to occlusion and given that in many scenes this occurs towards the bottom part of the body and along vertically-aligned locations, we also generate occluded examples using a small set of horizontal and vertical masks (fig. 5). This is, admittedly, an oversimplification but more complex combinations are possible. Ideally we would want to learn an occlusion prior for people in natural scenes. The database is somewhat artificial, but to our knowledge there is no realistic alternative that would contain images of humans with 3d ground truth.

The 'naturally enhanced' database is used to train the generative and the recognition model using maximum-likelihood [12, 11, 28]. The recognition model is a 5-component conditional Bayesian mixture of sparse linear experts. The linearity is important because during learning, sparsity constraints adaptively downgrade SIFT components and/or locations that are systematically found not to correlate with (*i.e.* not to be useful when predicting) the target subset of 3d poses allocated to different experts. The initial generative and recognition models are jointly refined using the algorithm given in fig. 2. In practice, this converges within 20-30 iterations. The learning steps usually inflate

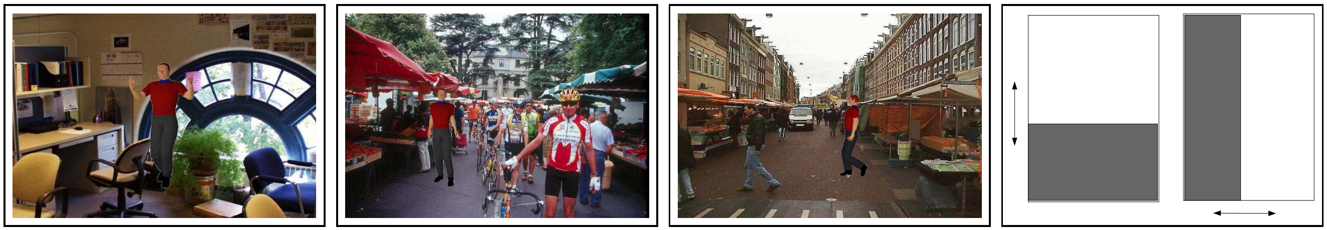

Figure 5. Examples of images included in our training database that consists of synthetically generated poses (from motion capture) rendered on natural indoor and outdoor image backgrounds. The rightmost plot show masks used to generate partial occlusion within the detection window (regions shown in gray) views. We use only simple combinations restricted to 2 horizontal and 3 vertical intermediate positions, but more complex arrangements are possible.

the covariance of the generative observation model, presumably to compensate for the inherent inaccuracy when averaging using the approximate recognition distribution Q_{ν}. This tends to make better use of the representation because additional samples, not originally in the supervised training set, are produced by the generative model and used to 'fill' the capacity of the recognition model.

The recognition model is used to detect 3d poses by scanning the image at different locations. For integrated detection and 3d reconstruction, we decide on the presence/absence of a human by training a classifier $f(g_1(\mathbf{r}), \ldots, g_M(\mathbf{r}), \mathbf{r})$ to predict the outcome. Besides the dependency on the input window descriptor, the classifier includes feedback from the generative model, *c.f*. (7) at the set of hypothesized 3d poses. This tends to make it overconservative but adding the generative residuals often increases performance.

Testing on artificial data with natural image background clutter: This series of tests is designed to evaluate the quality of our image descriptors under different operating conditions and two different models: one recognition model trained as before [28] and one trained using the algorithm described in §2. We have trained and tested both the HSIFT (table 1) and the BSIFT (table 2) in a variety of combinations, with and without background clutter (the training / testing runs without clutter are referred in the tables as 'clean', whereas the ones with clutter are referred as 'noisy'). In most tests, HSIFT has not produced accurate results. The recognition models we tried (various conditional mixtures of sparse linear and Gaussian kernel regressors) severely overfitted, having good performance on the training set and inaccurate one on the test set. This lack of generalization is observable under most training and testing combinations, except for clean backgrounds. The behavior is nevertheless to be expected – the histogram representation often smears the clutter in all the descriptor components.

The BSIFT descriptor we have tested behaved significantly better. This has been previously observed by [4] in pedestrian detection experiments. Differently however, we don't use overlapping blocks, because although they tend to slightly increase performance, they also increase the input descriptor size significantly, thus slowing down learning, which we do repeatedly. The results of our tests are given in table 2. The improved performance in clutter is caused by the positive influence of sparsity on the input entries that are systematically found not to correlate with (*i.e.* not be useful when predicting) the corresponding expert target 3d poses. The use of a generative model improves performance in most tests. It should be emphasized that although the gain may seem modest, an increase in error *per joint angle* of about $2-3^o$ (from a base of say, $5-6^o$) often crosses the border between the class of solutions that look qualitatively plausible, to poses that simply do not visually correlate with the image. Moreover, in our tests, the use of a generative model is limited to one-shot feedback, but doing more expensive mean field iterations (*c.f*. the E-step of fig. 2) will further improve performance.

	Recog	Recog + Gener
Clean train & test	5.8	4.9
Clean train, noisy test	18.3	17.2
Noisy train, clean test	19.2	18.2
Noisy train & test	15.7	15.1

Table 1. Comparative results (average error in degree per human joint angle) for training and testing different models using a 128d HSIFT: Recog is a feed-forward (recognition only) model. Recog + Gener is a combined recognition + generative model. We observe that this representation cannot generalize and the recognition model severely overfits. See also fig. 4.

Reconstruction results on real images: We have run several tests on outdoor images from a publicly available database [4] and on a sequence filmed in a laboratory. Some outdoor reconstruction results are shown in fig. 6. The images are quite difficult, because of self-occlusion and occasional lack of contrast. The recovered poses are not always accurate, with errors that concentrate in the elbow and shoulder complex. However, we feel that the results still capture the important aspects in the pose of the human subjects.

The other sequence we tried was filmed in a laboratory fig. 7, but under rather adverse conditions. These illustrate

	Recog	Recog + Gener
Clean train & test	5.6	3.9
Clean train, noisy test	10.2	7.9
Noisy train, clean test	7.3	5.8
Noisy train & test	8.2	6.7

Table 2. Comparative results (average error in degree per joint angle) for training and testing different models using a 8064d BSIFT descriptor collected densely inside the detection window: Recog is a feed-forward (recognition only) model. Recog + Gener is a combined recognition + generative model. This representation generalized better. See also table 1 and fig. 4.

some of the important difficulties of analyzing human motion in office environments: multiple lighting sources cast shadows that make background subtraction ineffective, several people may occlude eachother and may be occluded by other objects like chairs or office desks. It is precisely for these types of sequences that we have designed a training strategy based on partial, horizontally and vertically occluded views (*c.f.* fig. 5). However, we are currently not able to reconstruct poses when the detection window of different people overlaps by more than about 30%. But we are able to reconstruct 3d poses when the people undergo partial occlusion, especially when it is produced by somewhat homogeneous regions. We also find that both in this sequence and in the quasi-artificial tested ones, the generative feedback is helpful in downgrading symmetric ambiguities. For instance, in the sequence in fig. 7, one of the two people comes toward the camera and turns back. In this case, there are a number of competing pose solutions, separated by 180^o, that represent people who either face the camera or away from it. What seems to be highly distinctive in selecting the good solution are the features of the face (the eyes, nose or mouth). The feedback given by the generative model enhances this signal and downweights inconsistent configurations.

4. Conclusions

We have presented a framework to jointly learn a bidirectional generative-recognition model for 3d human motion reconstruction in monocular video. Our self-supervised learning algorithm alternates between training the recognition model using samples from the generative model and training the generative model to infer solutions close to the ones predicted by the recognition model. For fast expectation calculations, the recognition model is represented as a conditional mixture of mean field experts. On-line detection and reconstruction operate by scanning a window at multiple image locations and predicting 3d human poses using a recognition model with consistency feedback. Our experiments support the hypothesis that this strategy (in conjunction with noisy training) is promising for the automatic reconstruction of 3d human motion in monocular video sequences filmed in complex environments.

We currently study alternative more robust and informative image feature encodings. This includes feature selection and foreground-background enhancement methods centered around segmentation and basis pursuit. We also plan to augment the database with natural human poses either without ground-truth or with ground-truth provided by fitting generative models. Long term, we plan to investigate variational approximations for learning low-dimensional and multilayer models.

Acknowledgments: Cristian Sminchisescu has been partly supported by the NSF grant IIS-0535140. CS thanks Allan Jepson for insightful discussions and the anonymous reviewers for valuable comments.

References

[1] CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003. 5

[2] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV*, 2006. 2, 4, 5

[3] C. Bishop and M. Svensen. Bayesian mixtures of experts. In *UAI*, 2003. 3

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2, 5, 6, 8

[5] D. Demirdjian, L. Taycher, G. Shakhnarovich, K. Grauman, and T. Darrell. Avoiding the streetlight effect: tracking by exploring likelihood modes. In *ICCV*, 2005. 2

[6] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *CVPR*, 2000. 2, 4

[7] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR*, 2004. 2

[8] Z. Ghahramani and G. Hinton. Variational learning for switching state-space models. *Neural Computation*, 2000. 2

[9] G. Hinton, P. Dayan, B. Frey, and R. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 1995. 2

[10] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *ECCV*, 1998. 2, 4

[11] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *CVPR*, 2001. 2, 5

[12] M. Jordan. *Learning in graphical models*. MIT Press, 1998. 3, 4, 5

[13] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, (6):181–214, 1994. 3

[14] P. Kumar, P. Torr, and A. Zisserman. Learning layered motion segmentation of video. In *ICCV*, 2005. 2

[15] X. Lan and D. Huttenlocher. Beyond trees: common factor models for 2d human pose recovery. In *ICCV*, 2005. 2

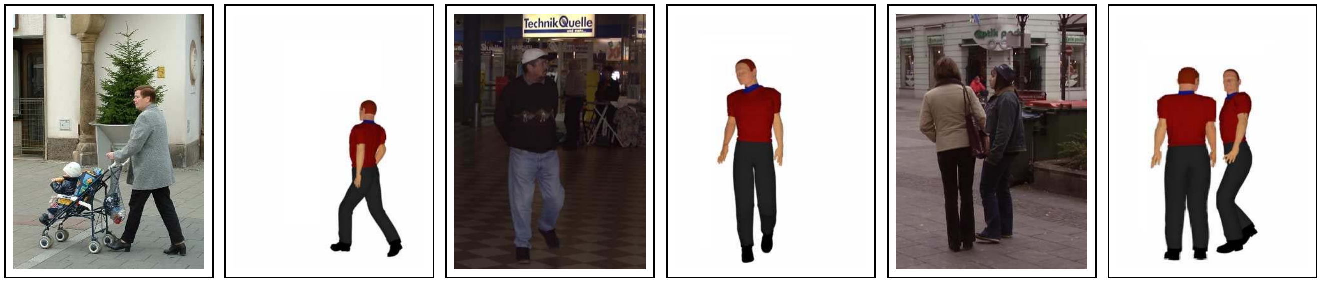

Figure 6. Several reconstruction from the database of [4]. Notice the difficulty of the poses that involve self-occlusion, clutter and sometimes low limb contrast. The position of the arms is not estimated precisely but the overall reconstruction still captures some of the important qualitative aspects in the posture of the human subjects.

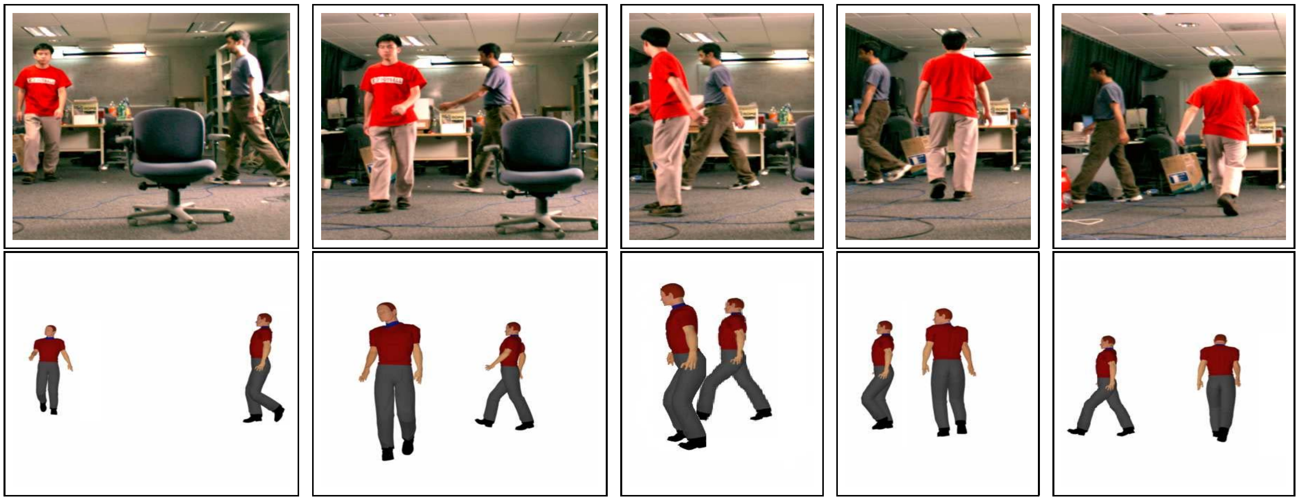

Figure 7. Reconstruction of two people in an office scene with occlusion and clutter. Notice the strong directional illumination sources that cast multiple shadows that make background subtraction ineffective. This sequence shows some of difficulties of tracking people in office spaces. The background is cluttered and there is occlusion from other objects (*e.g.* the chair) or people. Some of the recovered poses are not perfect, and we are currently not able to reconstruct poses when the detection windows of different people overlap by more than about 30%. But we are able to reconstruct under partial occlusion, especially when this comes from somewhat homogeneous regions.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of IEEE*, 1998. 4

[17] B. Leibe, E. Seeman, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005. 2

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 2, 4, 5

[19] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, 2004. 2

[20] V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A Dynamic Bayesian Approach to Figure Tracking using Learned Dynamical Models. In *ICCV*, 2001. 2

[21] D. Ramanan and C. Sminchisescu. Training Deformable Models for Localization. In *CVPR*, 2006. 2

[22] R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *NIPS*, 2002. 2, 3, 5

[23] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *ICCV*, 2003. 2, 5

[24] H. Sidenbladh and M. Black. Learning Image Statistics for Bayesian Tracking. In *ICCV*, 2001. 2, 4

[25] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking Loose-limbed People. In *CVPR*, 2004. 2

[26] C. Sminchisescu and A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *ICML*, pages 759–766, Banff, 2004. 2, 4

[27] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional Visual Tracking in Kernel Space. In *NIPS*, 2005. 4

[28] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, volume 1, pages 390–397, 2005. 2, 3, 4, 5, 6

[29] K. Toyama and A. Blake. Probabilistic Tracking in a Metric Space. In *ICCV*, 2001. 2

[30] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking in small training sets. In *ICCV*, 2005. 2

[31] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians using Patterns of Motion and Appearance. In *ICCV*, 2003. 2, 5