IEEE International Conference on Computer Vision (ICCV), November 2011, Barcelona, Spain.

Latent Structured Models for Human Pose Estimation

Catalin Ionescu[1,3], **Fuxin Li**[2], **Cristian Sminchisescu**[1,3]
[1]Faculty of Mathematics and Natural Sciences, University of Bonn
[2]Georgia Institute of Technology, [3]Institute for Mathematics of the Romanian Academy
{catalin.ionescu,fuxin.li,cristian.sminchisescu}@ins.uni-bonn.de

Abstract

We present an approach for automatic 3D human pose reconstruction from monocular images, based on a discriminative formulation with latent segmentation inputs. We advance the field of structured prediction and human pose reconstruction on several fronts. First, by working with a pool of figure-ground segment hypotheses, the prediction problem is formulated in terms of combined learning and inference over segment hypotheses and 3D human articular configurations. Besides constructing tractable formulations for the combined segment selection and pose estimation problem, we propose new augmented kernels that can better encode complex dependencies between output variables. Furthermore, we provide primal linear re-formulations based on Fourier kernel approximations, in order to scale-up the non-linear latent structured prediction methodology. The proposed models are shown to be competitive in the HumanEva benchmark and are also illustrated in a clip collected from a Hollywood movie, where the model can infer human poses from monocular images captured in complex environments.

1. Introduction

Reconstructing the tridimensional human pose and motion of people in the office, on the street, or outdoors based on images acquired with a single (or even multiple) video camera(s) is one of the open problems in computer vision. The difficulties compound: people have many degrees of freedom, deform and articulate, and their appearance varies widely as they span a significant range of body proportions and clothing. When analyzing people in realistic imaging conditions, the backgrounds cannot be controlled and people can also be occluded by other objects or people.

The 2D to 3D imaging relations and their ambiguities in the articulated case under point-wise joint correspondences are relatively well understood both in the monocular and in the multi-camera setting; several classes of methods exist: generative sampling strategies [8, 5], discriminative methods based on multi-valued predictors[1, 24], as well as methods that combine discriminative prediction and verification [20]. Techniques to reduce search complexity to low-dimensional non-linear manifolds have gained popularity[23, 25, 22], partly for efficiency considerations, but also as means to handle measurement ambiguities or missing data resulting from occlusion. Output manifold assumptions are just one way to model structure. More generally would be to use structured kernels[13].

Recently there is a trend towards operation in realistic environments where people are viewed against more complex backgrounds and in a more diverse set of poses. Handling such environments would be in principle possible by means of integrated systems that jointly combine person detection (or localization) and 3D reconstruction[13, 3, 2]. However human detectors[10] cannot always detect general human poses, and even when they do, figure-ground needs to be resolved from the bounding box before 3D pose can be predicted reliably (it is agreed that predictors based on silhouettes generalize relatively well if the input quality is good and the training distribution matches sufficiently well the set of human poses typical of the problem domain[1, 24, 20]). Another approach would be to use more detailed 2D human part-based models for localization[19, 2, 11, 9]. But 2D pictorial models usually come with a relatively high false positive rate and encounter difficulties localizing people viewed under sharp 3D effects (foreshortening). Closest to our goal of simultaneously obtaining segmentations and 3D poses is Posecut[4], a pose estimation method which alternates between fitting a 3D skeletal model to a silhouette (standard generative fitting or alignment), and re-estimating the silhouette (solving a binary MRF) using the predicted skeleton boundary as a shape prior.

In this paper we pursue a constrained discriminative approach that jointly estimates a quality selector function over a set of latent figure/ground segmentations (obtained using parametric max-flow and ranked based on mid-level properties) and a structured predictor. We introduce several novel elements that touch upon structural modeling, efficiency,

Figure 1. Illustration of our human localization and 3D pose reconstruction framework. Given an image (left), we extract a set of putative figure-ground segmentations by applying constraints at different locations and spatial scales in (monocular) images, using the Constrained Parametric Min-Cuts (CPMC) algorithm[6]. The method is tuned towards segments that exhibit generic mid-level object regularities (convexity, boundary continuity), but uses no person priors. Different segments extracted by CPMC are shown in images 2, 3 and 5,6 respectively. Given segment data together with 3D human pose information, we will jointly learn a latent structured model that knows both how to select relevant segments and how to predict accurate 3D human poses, conditioned on that input selection. Overall, we contribute with a latent formulation based on Kernel Dependency Estimation, with novel structured kernels that better encode correlations among the human body parts, and with Fourier kernel approximations that enable linear training for large datasets. Automatic 3D reconstructions obtained by our method are shown in images 4 and 7, with renderings based on synthetic graphics models.

and experimental realism.

First, a large set of figure-ground segments is generated for each image, at multiple locations and scales using Constrained Parametric Min-Cuts (CPMC), a recent segmentation procedure that has proven to be effective in a number of image segmentation[6] and labeling tasks[15]. We cast the problem of automatic localization and 3D human pose estimation as combined inference over segments and human articular configurations that give optimal prediction.

Second, we give a novel yet general and tractable formulation for latent structured models (with latent kernel dependency estimation l-KDE as a special case), as applicable to localization and continuous state estimation. We also propose augmented kernels for human pose and show that these are quantitatively better at representing complex interdependencies among the body parts than their unstructured counterparts. The computational cost of using accurate nonlinear kernels is one of the main challenges in scaling the methodology to large datasets.

Our third contribution is therefore the formulation of latent structured models like l-KDE or structured SVM[13] in the framework of random Fourier approximations[18, 26, 16]. We show that under a suitable change of variables, the calculations become linear, with primal formulations, as well as gradient calculations expressed directly in the linear space induced by the Fourier representation. We report quantitative studies in the HumanEva benchmark[21] and also show that our segmentation-based latent structured models produce promising 3D reconstruction results for people in a variety of complex poses, and filmed against more complicated backgrounds than previously.

2. Segmentation-based Pose Prediction

Our goal is to investigate the continuous structured prediction problem, with human pose estimation as a special case, under multiple, imperfect input hypotheses. Input segments that partly align with the person boundaries contain important cues for prediction. However, among imperfect segments, it is not obvious which measurement would be a good indicator of the usefulness of a segment for pose prediction. Segments with the same 70% (pixel union over intersection) overlap to the ground truth can be very different: some may miss limbs of the person, some may miss the head and some may cover the person and some background. Conceptually, pose estimation errors will differ among different imperfect segments, an observation confirmed by our experiments.

Because a clear-cut definition of segment quality remains elusive, in this paper we seek to learn a task specific function from data—one that quantifies whether a correct pose can be predicted from a segment. Input quality is in this respect latent, and can only be indirectly inferred by the extent an input is effective for prediction. We cast the joint segment selection and pose prediction problem as estimating two functions, a pose predictor $\mathbf{f}$ and a segment quality selector g. During testing time, the quality function g selects the most suitable segment, then $\mathbf{f}$ performs output prediction on this segment. Namely,

$$\mathbf{y} = \mathbf{f}(\mathbf{I}) = \mathbf{f}(\arg\max_h g(\mathbf{r}_h^{\mathbf{I}})) \tag{1}$$

where $\mathbf{y}$ is a D-dimensional vector of 3D relative joint positions, $\mathbf{I}$ can either be an image, or a region-of-interest in an image, assumed segmented into N figure-ground hypotheses $S^{\mathbf{I}} = \{\mathbf{r}_1^{\mathbf{I}}, \ldots, \mathbf{r}_N^{\mathbf{I}}\}$. By abuse of notation, we also use $\mathbf{r}_i^{\mathbf{I}}$ to represent the d dimensional descriptor extracted on the segment; $\mathbf{y}^I$ denotes the D dimensional ground truth pose for image $\mathbf{I}$ and s^I is the ground truth segment.

Learning is performed by optimizing $\mathbf{f}$ and g jointly. A conceptual formulation can be:

$$\begin{aligned} \min_{\mathbf{f},g} \quad & \sum_{\mathbf{I}} \sum_{h \in \mathbf{I}} g(\mathbf{r}_h^{\mathbf{I}}) \|\mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) - \mathbf{y}^{\mathbf{I}}\|^2 + \lambda_g \|g\|^2 + \lambda_{\mathbf{f}} \|\mathbf{f}\|^2 \\ \text{s.t.} \quad & g(\mathbf{r}_h^{\mathbf{I}}) \geq 0, \max_{h \in \mathbf{I}} g(\mathbf{r}_h^{\mathbf{I}}) \geq 1, \forall \mathbf{I} \end{aligned} \tag{2}$$

where $\|g\|$ and $\|\mathbf{f}\|$ are norms (*e.g.*, RKHS norms [12])

to prevent over-fitting. The formulation is similar in spirit to the Multiple Instance Learning Problem (e.g., [17]). Minimization of the objective function requires that $g(\mathbf{r}_h^{\mathbf{I}})\|\mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) - \mathbf{y}^{\mathbf{I}})\|^2$ is small. Therefore, the quality function g should give higher scores to segments that have low prediction error. To make the formulation practical, g needs to be always positive. Also, in order to avoid degeneracies *e.g.*, $g(\mathbf{r}_h^{\mathbf{I}}) \equiv 0$ at least one segment in each image must have a good score ($g(\mathbf{r}_h^{\mathbf{I}}) > 0$).

We take a logistic formulation and select $g(\mathbf{r}_h^{\mathbf{I}}) = \frac{1}{1+\exp(\mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))}$ to ensure positiveness. However it is not obvious how to set a convex constraint to avoid these degeneracies. For example a constraint like: $\sum_{h\in\mathbf{I}} g(\mathbf{r}_h^{\mathbf{I}}) \geq 1$ is not convex. We choose instead to introduce constraints on the sum: $\sum_{h\in\mathbf{I}} \mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}})$. Since a smaller $\mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}})$ implies a larger $g(\mathbf{r}_h^{\mathbf{I}})$, making the sum small makes some $g(\mathbf{r}_h^{\mathbf{I}})$ large. We use the binomial log-likelihood loss $L(x) = \log(1 + 2\exp(x))$ as a soft penalty on large values of $\sum_{h\in\mathbf{I}} \mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}})$. Putting it together, the optimization problem can be recast as follows:

$$\begin{aligned}\min_{\mathbf{f},\mathbf{w}_g} \quad & \sum_{\mathbf{I}} \sum_{h\in\mathbf{I}} \frac{1}{1+\exp(\mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))} \|\mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) - \mathbf{y}^{\mathbf{I}}\|^2 \\ & + \lambda_s \sum_{\mathbf{I}} \log(1 + 2\exp(\sum_{h\in\mathbf{I}} \mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))) \\ & + \lambda_g \|\mathbf{w}_g\|^2 + \lambda_{\mathbf{f}} \|\mathbf{f}\|^2 \end{aligned} \tag{3}$$

where λ_g, $\lambda_{\mathbf{f}}$ and λ_s are regularization parameters.

A quite subtle point in the optimization is the balance between the first and the second term in the objective. As argued before, the second term tends to drive $g(\mathbf{r}_h^{\mathbf{I}})$ large. But a large g value for all segments leads to large total weights and a higher weighted prediction error. Balancing the two terms in the optimization provides an appropriate quality function g which outputs large values for segments more suitable for prediction, and small values otherwise.

To solve the problem, we use alternating minimization on $\mathbf{f}$ and g (Algorithm 1). Estimation of $\mathbf{f}$ and g then can be done via any standard approaches. We have experimented with both kernel models and linear models based on random Fourier features (described in Sec. 2.2). The estimation of g is an unconstrained optimization which can be solved using an LBFGS algorithm provided, *e.g.* in the `minFunc` package[1]. The optimization for $\mathbf{f}$ is a weighted regression, as the second term does not take effect. More expensive options for learning $\mathbf{f}$, including a structured SVM model, are given in a companion technical report [14].

2.1. Kernel Dependency Estimation

Complex inference problems with interdependent inputs and outputs require models that can represent correlations both in the input and in the output explicitly. An efficient solution to model dependencies is kernel dependency estimation (KDE) where multi-dimensional continuous outputs are non-linearly decorrelated by kernel PCA. Prediction is then redirected to the kernel principal component subspace. Since in this new space dimensions are orthogonal, regression can be performed independently on each latent dimension. We use KDE as an alternative to independent-output regression for learning of $\mathbf{f}$. Based on (3) we consider a weighted KDE model with weights $g(\mathbf{r}_h^{\mathbf{I}})$ assigned to each segment

$$\min_{\mathbf{w_f}} \sum_{\mathbf{I}} \sum_{h\in\mathbf{I}} g(\mathbf{r}_h^{\mathbf{I}}) \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h^{\mathbf{I}}) - \Phi_Y(\mathbf{y}^{\mathbf{I}})\|^2 + \lambda_{\mathbf{f}} \|\mathbf{w_f}\|^2 \tag{4}$$

where $\Phi_X : \mathbb{R}^d \to \mathbb{R}^m$ is a non-linear map applied to the inputs and $\Phi_Y : \mathbb{R}^D \to \mathbb{R}^n$ is an orthogonal kernel PCA embedding of the targets.

Remarkably, the learning problem (4) can be solved in closed form [7] with parameters given by

$$\mathbf{w_f} = (\mathbf{M}_X^\top \Upsilon \mathbf{M}_X + \lambda_{\mathbf{f}} \mathbf{I})^{-1} \mathbf{M}_X^\top \Upsilon \mathbf{M}_Y \tag{5}$$

where Υ is the diagonal matrix with elements $g(\mathbf{r}_h^{\mathbf{I}})$, $\mathbf{M}_Y = [\Phi_Y(\mathbf{y}_1) \dots \Phi_Y(\mathbf{y}_N)]^\top$ the $N \times n$ matrix of outputs and similarly $\mathbf{M}_X$ the $N \times m$ matrix of input features.

Given an input and a model, the output is computed by solving for the pre-image of the projection in the original pose space. In this case we compute the minimum ℓ_2 distance between the reconstruction and the output in the latent (feature) space

$$\mathbf{y}_h^\star = \mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) = \arg\min_{\mathbf{y}} \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h^I) - \Phi_Y(\mathbf{y})\|^2 \tag{6}$$

where in testing the segment is selected based on (1).

The KDE regression framework is simple and efficient in both training and testing. Since no inference is performed during training, only n standard regressors (following kernel PCA on outputs) need to be estimated. The training is significantly faster than alternatives such as structural SVM (see our technical report [14]). Inference is also simpler

Algorithm 1 Algorithm for latent structured learning.

1: {Initialization of $\mathbf{f}$ using ground truth segments}
2: $\min_{\mathbf{f}} \sum_{\mathbf{I}} \|\mathbf{f}(s^{\mathbf{I}}) - \mathbf{y}^{\mathbf{I}}\|^2 + \lambda_f \|\mathbf{f}\|^2$
3: **while** not converged **do**
4: {Learn the segment ranker}
5: $\min_{\mathbf{w}_g} \sum_{\mathbf{I}} \sum_{h\in\mathbf{I}} \frac{1}{1+\exp(\mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))} \|\mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) - \mathbf{y}^{\mathbf{I}}\|^2 + \lambda_s \sum_{\mathbf{I}} \log(1+2\exp(\sum_{h\in\mathbf{I}} \mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))) + \lambda_g \|\mathbf{w}_g\|^2$
6: $g(\mathbf{r}_h^{\mathbf{I}}) = \frac{1}{1+\exp(\mathbf{w}_g^\top \Phi_g(\mathbf{r}_h^{\mathbf{I}}))}$
7: {Learn the pose predictor with the new weights}
8: $\min_{\mathbf{w_f}} \sum_{\mathbf{I}} \sum_{h\in\mathbf{I}} g(\mathbf{r}_h^{\mathbf{I}}) \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h^{\mathbf{I}}) - \Phi_Y(\mathbf{y}^{\mathbf{I}})\|^2 + \lambda_{\mathbf{f}} \|\mathbf{w_f}\|^2$
9: $\mathbf{f}(\mathbf{r}_h^{\mathbf{I}}) = \arg\min_{\mathbf{y}} \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h^I) - \Phi_Y(\mathbf{y})\|^2$
10: **end while**

[1] http://people.cs.ubc.ca/schmidtm/Software/minFunc.html

since the segment quality selector can be evaluated efficiently.

2.2. Random Fourier Approximations

Random Fourier approximations (RF) [18, 26, 16] provide an efficient methodology to create explicit feature transforms for non-linear kernel methods. In RF, explicit feature vectors are created for examples so that their inner products are Monte Carlo approximations of the kernel. This is important since linear methods typically scale linearly in the number of training examples, whereas kernel methods scale at least quadratically. Therefore, RF enables accurate training with large datasets in many circumstances.

The algorithm for the change of representation has two steps: i) Generate n random samples $\gamma_1, \ldots, \gamma_n$ from a distribution μ dependent on the Fourier transform of the kernel; ii) For all examples, compute the random projection [16]

$$\mathbf{Z}(\mathbf{u}) = \left[\cos(q_{\gamma_1}(\mathbf{u}) + b_1), \ldots, \cos(q_{\gamma_n}(\mathbf{u}) + b_n)\right]^\top \quad (7)$$

where $q_\gamma(\mathbf{u})$ is an inner product function depending on the kernel and b_i are uniform random samples drawn from $[0, 2\pi]$. The kernels used in this paper are the Gaussian kernel, where $q_\gamma(\mathbf{u}) = \gamma^\top \mathbf{u}$, for real-valued poses and the skewed chi-square kernel for nonnegative-valued histogram descriptors of the input [16], with $q_\gamma(\mathbf{u}) = \log(\gamma^\top \mathbf{u} + c)$, where c is the kernel parameter. For details see [18, 26, 16].

After applying the RF feature transform (7) learning is simply performed as linear least squares regression (or linear dependency estimation) on the RF feature matrix.

The RF methodology can be applied to our KDE model in a straightforward manner *i.e.* by using $\Phi_X = \mathbf{Z}_X$ and $\Phi_Y = \mathbf{Z}_Y$. Inference in the model remains non-convex, even for Fourier embeddings and we optimize locally using a BFGS quasi-Newton method. For optimization, the gradient of the inference function w.r.t. the inputs needs to be computed. The gradient in direction j, is obtained analytically by differentiating the feature map $\frac{\partial \mathbf{Z}_Y(\mathbf{y})}{\partial \mathbf{y}_j} = -\sum_{k=1}^{n} \frac{\partial q_{\gamma_k}(\mathbf{y})}{\partial \mathbf{y}_j} \sin(q_{\gamma_k}(\mathbf{y}) + b_k)$ to give:

$$\begin{aligned}
&\nabla_j \tfrac{1}{2} \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}) - \Phi_Y(\mathbf{y})\|^2 = \\
&= -\frac{\partial \mathbf{Z}_Y(\mathbf{y})}{\partial \mathbf{y}_j} (\mathbf{w_f}^\top \mathbf{Z}_X(\mathbf{r}) - \mathbf{Z}_Y(\mathbf{y})) = \\
&= \sum_{k=1}^{n} \frac{\partial q_{\gamma_k}(\mathbf{y})}{\partial \mathbf{y}_j} \sin(q_{\gamma_k}(\mathbf{y}) + b_k)(\mathbf{w_f}^{(k)} \mathbf{Z}_X(\mathbf{r}) - \mathbf{Z}_Y^{(k)}(\mathbf{y}))
\end{aligned} \quad (8)$$

We use superscripts (k), to denote the column of a matrix.

3. Augmented Output

In this section, our structured approach is extended to predict augmented outputs in addition to the pose vector. This helps regularize the result to valid poses. We draw from the intuition that for a valid human pose, the proportion of limb lengths are fixed. One straightforward approach would be to include these as constraints in the inference problem. However, when the pose is represented as 3D joint coordinates, one convex approach to enforce this constraint could be to use both l_1 distances and l_1 constraints, *e.g.*, $|\|\mathbf{y}_i - \mathbf{y}_j\|_1 - \|\mathbf{y}_l - \mathbf{y}_k\|_1| \leq L$. Inference with such non-smooth and complicated constraints is neither easy nor very elegant.

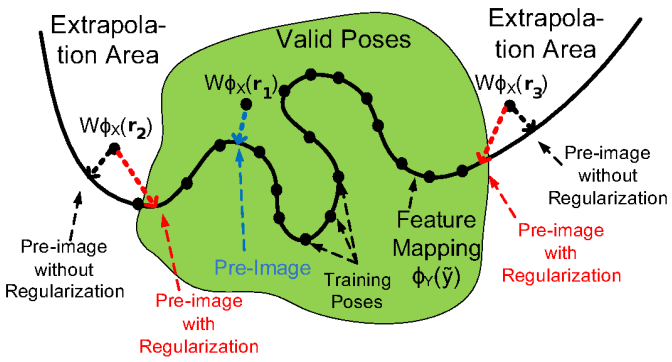

Figure 2. Illustration of the inference with the augmented kernel.

Instead, we propose to add auxiliary variables to the original output. The log limb length ratios (LLLR) $\mathbf{y}_a = \log \frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{V_{ijkl}\|\mathbf{y}_k - \mathbf{y}_l\|}$ of the relevant limbs are used as additional output dimensions, where V_{ijkl} are empirical average limb length ratios. Then kernel dependency estimation is performed on the combined output $\tilde{\mathbf{y}} = \left[\mathbf{y}, \mathbf{y}_a\right]$ consisting of 3D joint positions and auxiliary variables. During inference, the following problem is solved:

$$\arg\min_{\tilde{\mathbf{y}}, h} \|\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h) - \Phi_Y(\tilde{\mathbf{y}})\|^2 + \lambda \|\mathbf{y}_a\|^2 \quad (9)$$

Fig. 2 illustrates the need for regularization during inference. KDE embeds from an ambient output space to a high-dimensional space, where the original poses emerge as a manifold $\Phi_Y(\mathbf{y})$. However, the prediction $\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h)$ can be any point in space. Inference boils down to finding the closest pre-image in the ambient space, *i.e.*, map back to the manifold using the non-linear map $\Phi_Y(\mathbf{y})$. For instance, the projection of predicted $\mathbf{w_f}^\top \Phi_X(\mathbf{r}_h)$ is shown in dark blue. However, since the manifold extends to regions with no training examples (extrapolation areas), the pre-image calculation may also not produce a valid pose, *e.g.* $\mathbf{w_f}^\top \Phi_X(\mathbf{r}_2)$ and $\mathbf{w_f}^\top \Phi_X(\mathbf{r}_3)$ in fig. 2. The correlation between the augmented outputs and the pose is captured by the nonlinear transform Φ_Y (they will be coupled with any orthogonal basis generated from kernel PCA), thus regularization on the augmented outputs directly translates into regularization of the ambient pose variables. This biases the pre-image map to regions of valid outputs.

From the training pose data, we compiled all the possible limb length ratios and plot the mean and standard deviation in fig. 3. It can be seen that many limb pairs have fixed ratios that do not change much across different subjects. We

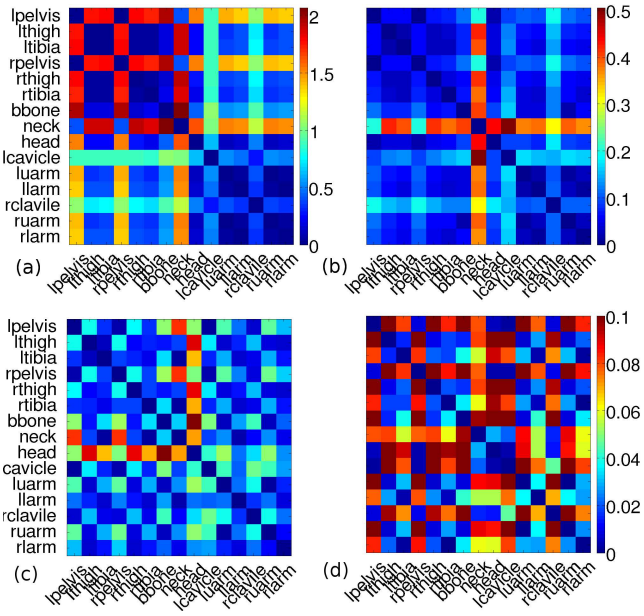

Figure 3. Empirical measures for the log of limb length ratio (LLLR) on the test set ground truth data show structure: mean (a) and standard deviation (b) . In the lower row we show l_1 error of same measure (LLLR) between predictions and the ground truth test data with KDEa (augmented) (c) and KDE (standard) model (d) (same color scale). KDEa predicts ratios closer to the ground truth, which partly explains its better overall performance.

choose such consistent limb ratios, and predict a 162 output $\mathbf{y}_a$ as constraint on the pose. The augmented output approach is not confined to a 3D representation of the pose based on joint positions. If joint angles need to be predicted, physical angle limits or other angle constraints (*e.g.* directional preference) can be included.

4. Experiments

We run experiments in support of our three main contributions. First, we show that coupled latent segment selection and pose prediction performs better than two independent classification and pose estimation stages. Second, we show that a structured model like KDE boosts pose prediction performance especially in conjunction with improved augmented kernels. Finally, we show that random Fourier techniques extend to a structured learning framework with nontrivial inference like the pose prediction, at no significant loss in performance and with the added benefit of scalability.

Dataset and Image Processing. We present a set of comprehensive experiments in the HumanEva-I dataset[21]. The dataset contains 5 motions (Box, Gestures, Jog, ThrowCatch and Walking) from 4 subjects. Accurate image and pose data (3D joint locations) are available for 3 subjects and 4 motions (we discard ThrowCatch since ground truth

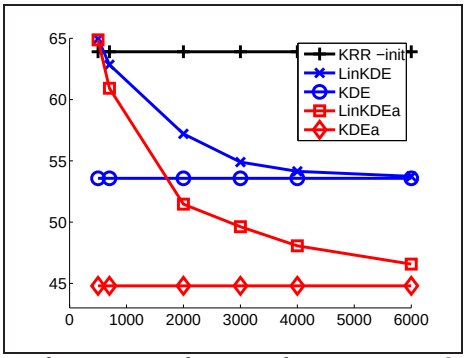

Figure 4. Performance as a function of approximation dimensionality in input space. The output kernel of the KDE is kept fixed throughout the experiment at 2000 dimensions while the augmented kernel is approximated with 4000 dimensions.

poses could not be accurately obtained). Together, the information from the 3 color cameras, offers a relatively large number of images and associated 3D pose data that can be used for training and testing—in total around 34,000 examples (divided roughly equally in a train and a test set). Ground truth segments to evaluate predictor baselines are obtained from background subtraction using the original code provided with the dataset. We additionally remove shadows in order to obtain accurate bounding boxes and segments. Note that we do not exclude data where background subtraction cannot be perfectly obtained, we only discard invalid pose data. Performance on this dataset is measured, in the standard way, as mean 3D joint errors, in mm.

Our method requires generating a pool of segments for which we use the CPMC algorithm [6] with standard parameters. Segmentation on this dataset was quite successful with a mean overlap (intersection over union) of the best segment at 72%. The image features used are pyramid block SIFT with 3 levels (2x2, 4x4 and 8x8 cells) with 9 gradient bin orientations (0-180 degrees, unsigned) extracted on the segment support in the image (not just its boundaries). These features are sensitive to the accuracy of the bounding box enclosing the segment so morphological operations were applied to remove very thin, protruding regions from segments.

Pose Prediction on Ground-truth Inputs. To calibrate the baseline performance of our regressor, we first show results on both ground truth segments as well as CPMC's best overlapping segment (table 1). In all experiments that illustrate KDE, KDEa, LinKDE and LinKDEa for inference, initialization was performed using the same Kernel Ridge Regression model (KRR) trained to predict outputs independently. It can be seen that KDE improves over KRR but it is also clear that it doesn't quite manage to get the proportions of the limbs right. Notice the boost in performance achieved by KDEa when the constraints were enforced. Figure 3 shows, on the first column, the distribution of the limb

Motion	KRR	KDE	KDEa	LinKDE	LinKDEa
Box	91.36	82.50	75.95	82.88	76.33
Gestures	68.96	66.89	60.24	65.35	61.70
Jog	62.09	50.26	44.92	51.64	46.54
Walking	63.75	52.40	44.18	53.57	46.41
All motions	70.11	65.50	61.28	67.79	63.70

KRR	KDE	KDEa	LinKDE	LinKDEa
86.53	78.60	71.77	78.92	70.36
81.69	76.05	71.56	76.81	72.17
54.17	45.73	40.16	45.93	41.13
55.27	44.48	37.73	45.66	38.79
65.24	63.93	58.48	63.03	59.90

Table 1. Mean per joint error (in mm) for different pose prediction models on HumanEva I data subjects 1-3 and data from all color cameras. Both activity specific models and activity independent models are reported. The left table represents the results on the segment with best overlap with ground truth for both training and testing. The right table shows the same results on ground truth segments for comparison.

ratios, and, on the second, the difference between the limb ratios obtained from predictions based on KDEa and KDE, with respect to the ground truth.

By applying the random Fourier methodology to structured models like KDE, KDEa with 400 and 6000 Fourier embedding dimensions on inputs we obtain a performance comparable with the kernel counterparts (with the same parameters). Notice however that linear methods scale much better in the size of the dataset, both in training and in testing. We see that indeed, even for complex models like KDE and KDEa the approximation holds well (fig. 4).

Combined Automatic Model. In order to assess the effectiveness of our segment selection procedure we have devised a set of baselines. The experiments were conducted on the best 5 segments based on overlap. Our aim was (*i*) to see if overlap is a good measure for selecting segments if the final goal is pose prediction and (*ii*) to assess the performance of our selection procedure. Results are shown in table 2. The first two columns show the 3D pose reconstruction error using the segment selection model described in section 2.1, with LinKDE and LinKRR for pose prediction. In this experiment we use only 2000 Fourier features for the input approximation. This explains the somewhat lower performance compared to our previous results. The initialization of the pose prediction models (1st step of the algorithm) is done using the ground truth segments. We have considered initialization using all segments, with the ground truth overlap as a segment ranking function $g(\mathbf{r}_h^I)$, but found that to work less well. Therefore the rest of the results shown in table 2 are all using LinKDE for pose prediction.

To assess the performance of a purely detection-based approach to segment selection we use the best overlap to the ground truth as segment selector. This is a sensible candidate for comparisons with a detection method since it is the perfect overlap-based selector. The third column in Table 2 shows the performance of the pose prediction model, tested on the segments with best overlap to the ground truth. The relatively low quality result in most cases means overlap is not necessarily a good selection criterion if one requires accurate pose prediction. In general, we usually only have segments with ground truth overlap of around 70% even in controlled environments. The results show that the segment having the very best overlap score was not the most informative for pose.

Motion	LinKRR	LinKDE	Overlap	Best	Mean
Box	97.44	84.61	105.31	69.12	84.59
Gestures	65.17	60.42	101.99	53.01	61.56
Jog	64.40	52.46	61.51	42.39	55.32
Walking	71.82	56.12	67.97	44.67	58.10

Table 2. Pose prediction error results (in mm) for the combined model for segment selection and pose estimation. We also show a simple LinKRR model baseline. LinKDE is the combined model based on linear Fourier embeddings. Third column (Overlap) gives pose prediction results based on the ground truth overlap selection. 'Best' is the highest accuracy results result that can be achieved by the LinKDE model if the ground truth pose were known. 'Mean' gives another baseline scenario where the best 5 segments predict the output by equal voting.

The 4th column of table 2 represents the best result we can achieve. This is computed by using the ground truth pose to select segments giving the lowest pose prediction error. The good results in this column indicate that our weighted KDE model is well trained and a better selection procedure could improve it further. The last column shows the mean error, for a model where the best 5 segments predict the output by equal voting.

Computational Efficiency. We approximate two different kernels, one over the inputs and the other one over outputs. The effect of this approximation is to transform a complexity that is quadratic in the size of the training set into a linear one. Training in this case consists of two steps: computation of the Fourier features for both inputs (of dimensionality m) and outputs (of dimensionality n) and solving a regression problem between the two. Let N be the number of training examples, $N \gg m$ and $N \gg n$. The complexity of LinKDE is $\mathcal{O}(Nm) + \mathcal{O}(m^2n + mn^2)$ (we consider matrix inversion to have quadratic complexity for simplicity). Contrast this with $\mathcal{O}(N^2)$ for the standard kernel method. For training sets beyond 10,000 examples, matrix inversion becomes difficult to perform. Moreover, for the combined segment selection and pose estimation model presented in §2 of the paper, our model increases 5 fold, since 5 segments are considered per image. The matrix inversion for our random Fourier formulation is independent of N (though con-

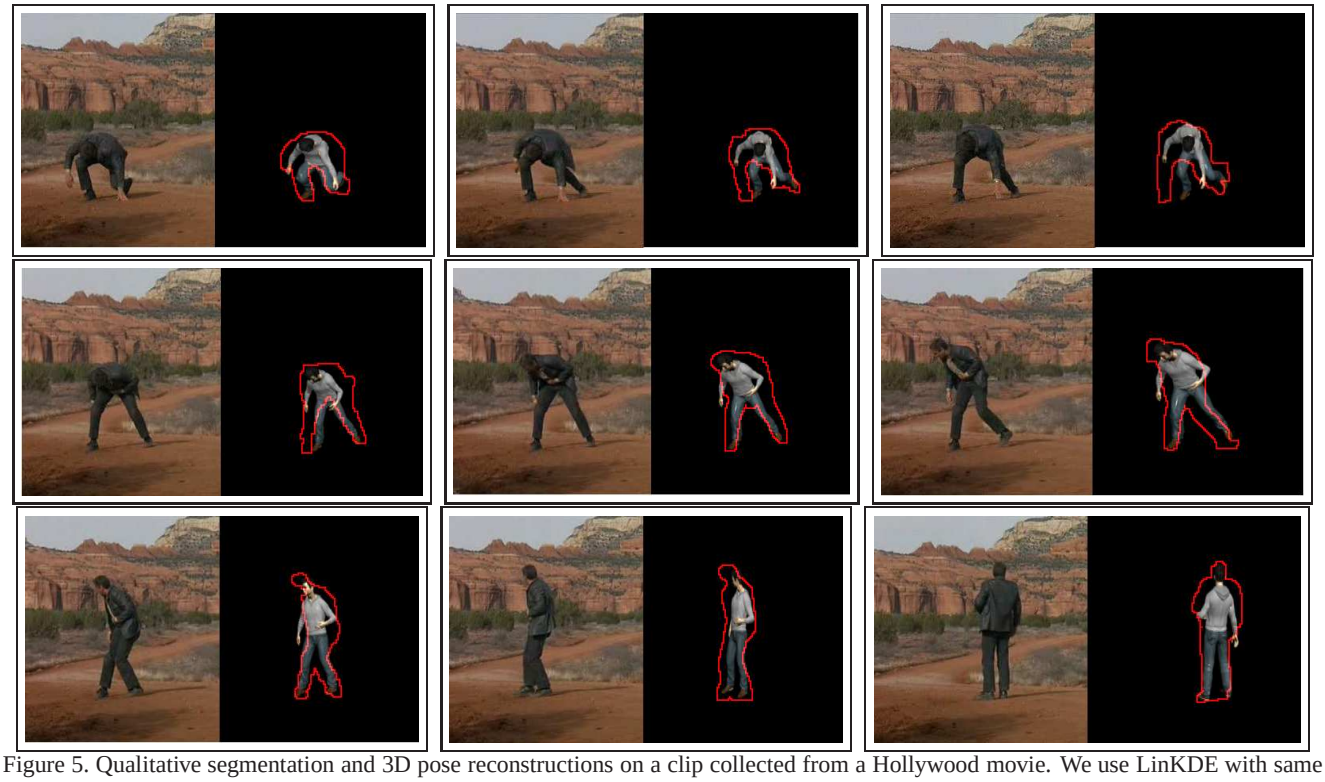

Figure 5. Qualitative segmentation and 3D pose reconstructions on a clip collected from a Hollywood movie. We use LinKDE with same input features and parameters as in the HumanEva evaluation, but with joint angle outputs. Using our Mocap system we have created a training set of 4000 examples, out of which 450 were qualitatively similar to the ones in the video and the remaining ones were different sitting and standing poses. The purpose of this experiment is to demonstrate the potential of the method in an un-instrumented environment and for more challenging poses.

structing the matrix is a linear operation in the number of examples).

In an experiment shown in table 2, we assess the impact of the input kernel approximation (here we fix the output approximation to 2000 dimensions) on training and testing times. Perhaps surprisingly, the testing time decreases with the dimensionality of the input approximation, an effect that can be explained by the increased gradient accuracy for higher-dimensional models. This is also true for the kernel version but the function evaluation is more costly involving the input matrix which is $\mathcal{O}(N^2)$.

We also study the impact of the output approximation dimensionality on training and testing times. Table 4 shows our results for a dataset of walking motions, and using the best segments with respect to overlap to the ground truth. Input dimensionality is fixed throughout the experiment at 4000. Training time is almost independent of the output approximation. Inference time however is affected, and roughly doubles when going from 500 to 6000 dimensions. Inference in the deterministic model is fast, but notice the small training set, which we use in order to be able to also perform exact calculations. The trend clearly reverses for larger datasets where above a certain size exact calculations become unfeasible.

5. Conclusion

We have presented a segmentation-based framework for automatic 3D human pose reconstruction in monocular images, based on a latent variable formulation for person localization in the image (by selecting over multiple figure-ground hypotheses) and 3D articular pose prediction. Learning the latent model is formulated as an alternating optimization. We give new formulations for latent kernel dependency estimation and show that such a methodology can be made scalable while preserving accuracy by means of linear Fourier approximations. We also introduce augmented structured kernels to improve the quality of output prediction. In extensive quantitative experiments we demonstrate that our model can jointly select accurate segments and provides promising automatic pose prediction results in the HumanEva benchmark. We also show results in a clip collected from a Hollywood movie, where more complex human poses were reconstructed. The augmented output provides a practical approach to incorporate pose constraints. The force one needs to exert to maintain a pose can also be computed with a physical model [5] and used as a prior for pose prediction. Such constraints have good potential to improve realism and will be investigated in future work.

Model	500	700	1000	3000	4000	6000	KDE
LinKDE -train	13	16	28	409	782	3167	109028
LinKDE -test	5739	5996	5574	5441	3831	2244	8811
LinKDE -test acc	73.56	73.03	71.99	68.28	67.32	66.58	65.57

Table 3. Running time as a function of approximation dimensionality of the *input* kernel (in seconds). The results are obtained on the entire dataset (all motions), with computations run on a Xeon 1 core with 48GB RAM. Clearly training time is prohibitive using the deterministic method but manageable even for high approximation dimensions using the Fourier methodology. The observed decrease in computation time can be caused by a more accurate gradient which makes the optimization more stable.

Model	500	700	1000	3000	5000	6000	KDE/KDEa
LinKDE -train	614.53	619.11	623.34	630.18	628.07	634.02	2219.12
LinKDE -test	285.22	288.13	289.37	437.95	613.03	698.71	642.36
LinKDEa -train	1524.99	1521.11	1520.84	1529.99	1528.16	1512.87	2224.12
LinKDEa -test	2409.01	2197.47	2525.41	2256.02	2782.98	3021.54	2444.29

Table 4. Computation time results (in seconds) for training and testing using our two LinKDE models with different number of random Fourier features approximating the *output* kernel. In the rightmost column we show the performance of standard KDE. Results are reported for a 4K dataset of walking motions, with calculations performed on a quad core Pentium processor. For LinKDEa, the kernel is computed as a sum of 2 approximations, where the approximation of the augmented component has 2000 Fourier dimensions. The input kernel was approximated using 4000 Fourier dimensions.

Acknowledgements: This work was supported by CNCSIS-UEFISCDI, under PNII-RU-RC-2/2009, and by the EC, MCEXT-025481. We thank Dragos Papava at IMAR for support with visualization and motion capture.

References

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

[2] M. Andriluka, S. Roth, and B. Schiele. People Tracking-by-Detection and People-Detection-by-Tracking. In *CVPR*, 2008.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[4] M. Bray, P. Kohli, and P. Torr. Posecut: Simultaneous segmentation and and 3d pose estimation of humans using dynamic graph cuts. In *ECCV*, 2006.

[5] M. Brubaker and D. Fleet. The Kneed Walker for Human Pose Tracking. In *CVPR*, 2008.

[6] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*, 2010.

[7] C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *ICML*, pages 153–160, New York, NY, USA, 2005. ACM.

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

[9] M. Eichner and V. Ferrari. We are family: Joint Pose Estimation of Multiple Persons. In *ECCV*, 2010.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32:1627–1645, 2010.

[11] V. Ferrari, M. Marin, and A. Zisserman. Pose Seach: retrieving people using their pose. In *CVPR*, 2009.

[12] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning, Jan 2008.

[13] C. Ionescu, L. Bo, and C. Sminchisescu. Structural SVM for Visual Localization and Continuous State Estimation. In *ICCV*, 2009.

[14] C. Ionescu, F. Li, and C. Sminchisescu. Fourier structured learning. Technical report, INS, University of Bonn, 2011.

[15] F. Li, J. Carreira, and C. Sminchisescu. Object Recognition as Ranking Holistic Figure-Ground Hypotheses. In *CVPR*, 2010.

[16] F. Li, C. Ionescu, and C. Sminchisescu. Random Fourier approximations for skewed multiplicative histogram kernels. In *LNCS (DAGM)*, September 2010.

[17] F. Li and C. Sminchisescu. Convex Multiple Instance Learning by Estimating Likelihood Ratio. In *NIPS*, 2010.

[18] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007.

[19] B. Sapp, A. Toshev, and B. Taskar. Cascaded Models for Articulated Pose Estimation. In *ECCV*, 2010.

[20] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007.

[21] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. volume 1, 2006.

[22] L. Sigal, R. Memisevic, and D. Fleet. Shared Kernel Information Embedding for Discriminative Inference. In *CVPR*, 2010.

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

[24] C. Sminchisescu, A. Kanaujia, and D. Metaxas. BM^3E: Discriminative Density Propagation for Visual Tracking. *PAMI*, 2007.

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

[26] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.