Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation

Catalin Ionescu[†♮], **Joao Carreira**[‡], **Cristian Sminchisescu**[♯†]

[♯]Department of Mathematics, Faculty of Engineering, Lund University [♮]University of Bonn
[†]Institute of Mathematics of the Romanian Academy [‡]University of California at Berkeley
catalin.ionescu@ins.uni-bonn.de, carreira@eecs.berkeley.edu, cristian.sminchisescu@math.lth.se

Abstract

Recently, the emergence of Kinect systems has demonstrated the benefits of predicting an intermediate body part labeling for 3D human pose estimation, in conjunction with RGB-D imagery. The availability of depth information plays a critical role, so an important question is whether a similar representation can be developed with sufficient robustness in order to estimate 3D pose from RGB images. This paper provides evidence for a positive answer, by leveraging (a) 2D human body part labeling in images, (b) second-order label-sensitive pooling over dynamically computed regions resulting from a hierarchical decomposition of the body, and (c) iterative structured-output modeling to contextualize the process based on 3D pose estimates. For robustness and generalization, we take advantage of a recent large-scale 3D human motion capture dataset, Human3.6M[18] that also has human body part labeling annotations available with images. We provide extensive experimental studies where alternative intermediate representations are compared and report a substantial 33% error reduction *over competitive discriminative baselines that regress 3D human pose against global HOG features.*

1. Introduction

We focus on the inference of 3D human pose from monocular intensity images (RGB). Recently, Kinect systems[16] based on discriminative learning methods and large-scale datasets of synthetically generated depth maps with dense body part annotations have achieved considerable success for 3D human pose prediction indoors. An open question is whether similar advances would be possible for 3D human pose estimation based on standard intensity images. The transition is by no means trivial, as depth allows the design of distinctive, scale and illumination-invariant local features and substantially constrains 3D inference. For RGB, 3D needs to be aggregated from 2D cues only, and 3D inference ambiguities can occur[32, 30].

Besides heavily relying on depth sensors, a novel aspect of Kinect was the use of an intermediate 2D part labeling stage within a 3D pose estimation pipeline trained on unprecedentedly large human pose datasets. No publicly available datasets have been available to match the diversity and size of the proprietary ones used for learning Kinect models so far. However, a large data gathering effort has resulted in the recently released Human3.6M dataset[18]. The annotations provided with this data include not only 3.6 million ground truth 3D pose and camera parameter configurations, figure-ground segmentation masks, and human body scans, but also automatically generated pixel-level body part labels, obtained using a volumetric human model, and the kinematic information available in the dataset. We will use the data to train human body labeling and 3D pose prediction models in multiple passes.

Specifically, our formulation decomposes the problem into three stages: (1) dense 2D human body part labeling, (2) *second-order label-sensitive pooling* over a hierarchical decomposition of the body, and (3) human 3D pose prediction by regressing against pooled descriptors, dynamically contextualized by pose estimates. The success of such a pipeline fundamentally depends on being able to label body parts in intensity images, and on the way such inherently noisy inferences are further processed. We target robust descriptions based on body part labeling using two complementary strategies. On one hand we design partially redundant global pose descriptors that operate on overlapping regions resulting from a hierarchical decomposition of the human body. We generalize recent region descriptors for semantic segmentation under high-order statistics[7] towards a novel description process, called *second-order label-sensitive pooling (O_2LP)*, where regions are identified dynamically based on their inferred labels. This allows us to derive efficient descriptors, with multiple levels of selectivity and invariance, for robust human pose estimation.

Additionally, we iteratively inject 3D pose information into the labeling process, thereby promoting globally consistent estimates. For example, the leg on the right side of a body in the image may be a right leg or left leg depending on whether the person is facing the camera or not. Using a 3D pose estimate as additional feature to contextualize labeling allows to transfer information between distant body parts, reducing ambiguity.

We conclude with experimental studies and a quantitative analysis of the proposed components, showing that our methodology outperforms state of the art discriminative baselines that regress 3D pose against global HOG silhouette features, by a substantial 33% margin. Fig. 1 illustrates the proposed model.

1.1. Related work

Many of the early approaches to 3D human pose estimation employed generative models to search the state space in order to identify configurations of a body model that would best align with image features[12, 32, 29, 34, 15]. While powerful, these models required careful initialization, considerable manual effort to design a realistic synthetic model and were computationally expensive. The problem of joint segmentation and pose prediction has also been attacked using top-down[6] and bottom-up[17] approaches. Discriminative regression methods[1, 31, 28, 20, 17] focus on prediction from image descriptors, but depend on the existence of a sufficiently large and diverse training set, which until recently has not been available. The design of sufficiently stable and accurate descriptors is also a main challenge. For example, the commonly adopted HOG filters[10] use a fixed grid of gradient histograms. Depending on the pose of the person in the image, the same body parts are likely to fall into very different cells. Because the input descriptor is unstable across poses, the mapping between the feature space and the pose space is irregular and complex to learn. While hierarchical encodings computed over regular grids have shown a degree of invariance to such factors in 3D human pose estimation[20], they remain pose-independent. An alternative would be to compute 'cells' that are pose-centric and can be obtained, *e.g.*, from an image estimate of the 2D body part layout.

Here we pursue such a representation in the form of a dense 2D body part labeling. Kinect[16] used fast feature extraction based on depth differences inspired by [23], to infer a dense part labeling of the human body from depth images. Body labeling methods have also been proposed in the context of person detection[4], to learn a model of the body[13] and recently for 2D pose prediction[21, 11], but not within models and algorithms carrying all the way to 3D pose estimation. These models handle spatial label dependencies using CRF formulations, which makes inference hard in loopy graphs. Others have recently proposed to handle complex dependencies by means of a sequential feedback mechanism, such as auto-context[33] and fixed-point structured labeling[25, 3]. Such approaches classify a pixel using context, based on previously predicted labels of neighboring pixels. They differ from us in that auto-context learns a sequence of models where earlier ones of the same type provide context for later ones. Fixed-point structured labeling[25] learns a single model that feeds back outputs into itself and is learned using both clean and purposively corrupted ground truth context data. Our model is somewhat related to auto-context but we use a deeper, higher-level form of context, a 3D pose estimate, as well as different description methods based on label-sensitive pooling.

There is a large body of literature on 2D pose prediction, with the dominant approaches employing pictorial structures[14, 27, 2, 35], which impose tree-structured dependencies between joint positions while maximizing an appearance likelihood. Body part labeling at the level of pixels operates at a different granularity. While the labeling of each individual pixel is performed under weaker spatial constraints, the average process could be more robust to individual component errors (*e.g.* estimating a pixel, but not an entire body part incorrectly). This appears adequate when pooling inherently noisy 2D label estimates for 3D pose prediction. It also has the potential advantage of labeling arbitrary partial views of the person (a property also shared by poselets[5], at their different level of bounding box granularity), which would be more challenging for pictorial structures, where a spatial dependency among a set of components, assumed visible, is sought. Ultimately, such complementary methodologies may have a role, at different levels, in a robust system in the long run.

2. Overview and Model Formulation

Our pose estimation methodology is based on descriptors computed from body part labels with an iterative scheme. The model we propose can be viewed as two coupled predictors: a human body part labeling method and a 3D pose predictor operating on features extracted based on a 2D labeling process.

Let $\mathcal{L}$ be a discrete set of labels and $\mathcal{P}(\mathcal{L})$ its powerset. Let $\mathbf{I}$ be an image consisting of N pixels, and $\mathbf{x}$ a vector, containing the label at each pixel in the image *i.e.* $\mathbf{x} = \{x_k \in \mathcal{L} | k = 1, \ldots, N\}$. Let $\mathbf{z_I} \in \mathbb{R}^{3 \times D}$ be a vector of 3-dimensional positions, for a representation with D joints, corresponding to the pose of the person in image $\mathbf{I}$[1].

Let $\mathcal{R} : \{1, \ldots, M\} \to \mathcal{P}(\mathcal{L})$ be a mapping, not necessarily onto, defining the assignment of M regions to elements of the power set of labels $\mathcal{P}(\mathcal{L})$. Let region $R_j = \{k | \mathbf{x}_k \in \mathcal{R}(j)\}$ be defined as the set of pixels with labels

[1] Whenever not critical, the dependence of different variables on $\mathbf{I}$ will be dropped.

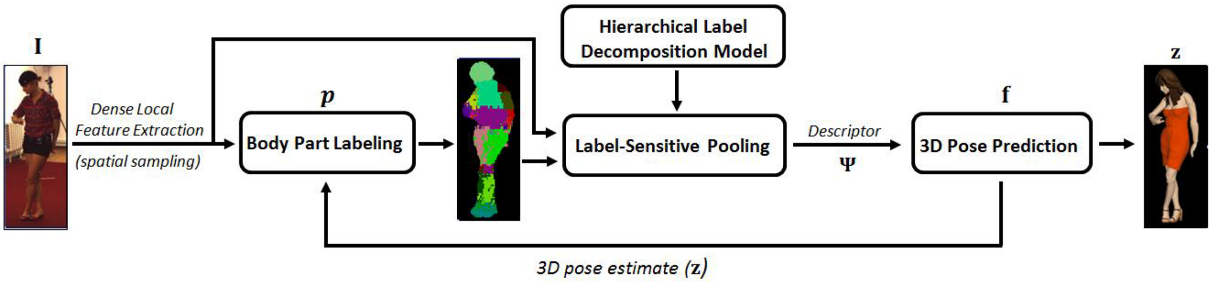

Figure 1: Overview of our 3D human pose estimation methodology. Local descriptors are extracted densely over the image and given as input to a 2D body part labeling model. A second order pooling process informed by a hierarchical decomposition of the body (individual limbs, upper and lower body parts, full body, *etc.*) is used to construct a global representation by stacking descriptors computed over overlapping subsets. 3D pose is obtained using regression, with estimates fed-back in order to further constrain the 2D labeling process.

in the set $\mathcal{R}(j)$. Let $\mathbf{s}(\mathbf{I}) = \{\mathbf{s}_k | k = 1, \ldots, N\}$, with $\dim(\mathbf{s}_k) = S$, descriptors extracted over neighborhoods centered at pixel k (*e.g.* SIFT) in $\mathbf{I}$. Let the statistical model producing the labeling, at iteration i be $p^{(i)}$ and the pose predictor at the same iteration be $\mathbf{f}^{(i)}$. Let $\mathbf{\Psi}(\mathbf{x}^{(i)}, \mathbf{I})$ be the pooled feature vector computed based on the labeling produced at iteration i. The model can be written as:

$$\mathbf{x}^{(0)} \leftarrow \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{I}) \tag{1}$$

$$\mathbf{z}^{(i)} \leftarrow \mathbf{f}^{(i)}(\mathbf{\Psi}(\mathbf{x}^{(i)}, \mathbf{I})), \forall i \geq 0 \tag{2}$$

$$\mathbf{x}^{(i)} \leftarrow \arg\max_{\mathbf{x}} p^{(i)}(\mathbf{x}|\mathbf{z}^{(i-1)}, \mathbf{I}), \forall i \geq 1 \tag{3}$$

We train by alternation, initially setting up an estimator for the labeling based on the image information alone (iteration 0), as $p^{(0)} \equiv p(\mathbf{x}|\mathbf{I})$, then we iterate between training a 3D pose model based on the current label predictions and training a labeling model using predicted poses as global context. At iteration i, $p^{(i)} \equiv p(\mathbf{x}|\mathbf{z}^{(i-1)}, \mathbf{I})$. In training we have to learn a set of models that use noisy inputs as pose context. In testing, we iterate using the trained models, $(p^{(i)}, \mathbf{f}^{(i)})$. The scheme usually converges within a few steps (see fig.3).[2]

3. Dense 2D Human Body Part Labeling (p)

We use multi-class random decision forests for body part labeling. The forest is a collection of T decision trees constructed over features φ, locally extracted in the image:

$$p(\mathbf{x}|\mathbf{I}) = \frac{1}{T} \sum_{t=1}^{T} p_t(\mathbf{x}|\mathbf{I}) \tag{4}$$

The trees contain split nodes and leaf nodes. Each split is associated to a component of the feature vector $\boldsymbol{\varphi}$ and a threshold τ. To label the pixel $\mathbf{x}_k$, we extract its corresponding feature vector $\boldsymbol{\varphi}_k$, start at the root and evaluate the feature components based on the corresponding thresholds. By reaching a leaf node, we can use the empirical distribution of stored labels (a majority vote or any other learnt model) in order to make the labeling decision.

In our formulation, the feature vectors over which decision trees are constructed are the previously introduced local S-dimensional vectors $\boldsymbol{\varphi}(\mathbf{x}, \mathbf{I}) \equiv \mathbf{s}$, for the labeling model $p(\mathbf{x}|\mathbf{I})$ and the $(S + 3 \times D)$-dimensional vectors $\boldsymbol{\varphi}^{(i)}(\mathbf{x}, \mathbf{I}) = (\mathbf{s}^\top, {\mathbf{z}^{(i-1)}}^\top)^\top$ for the contextual model $p(\mathbf{x}|\mathbf{z}^{(i-1)}, \mathbf{I})$. For the latter, the vector $\mathbf{z}$ is obtained as the pose estimate of the model $\mathbf{f}$, at the previous iteration $i - 1$. Contextual features are computed at different pixel locations k for each image $\mathbf{I}$, and will include a global pose component, estimated for that image. The local descriptors will not change, but the context will change, leading to random forests that depend on the iteration index.[3]

Features and Encodings: Different features $\boldsymbol{\varphi}$ were used by the labeling models we developed. The simplest one described the local image patches $\mathbf{s}_i$ using SIFT (the RF method). Our experiments (table 1) however showed that a model based on plain SIFT did not perform very well. During analysis, we identified 3 problems: (1) inherent ambiguity between limbs because locally arms and legs look similar, (2) lack of repeatability due to descriptors computed for images where people appear at different scales, and (3) ambiguities between the left and right sides of the body.

We address (1) by adding the location of the pixel relative to the person bounding box coordinate system as an additional feature (lRF) to $\boldsymbol{\varphi}$. (2) is addressed by learning a scale predictor based on simple features and using the es-

[2] Structured prediction methods like SOAR[3] offer a different approach to iteratively feed the output (pose) estimate as a covariate together with the image descriptor, although their analytical formulation would not immediately extend to a dynamic, pose-dependent descriptor construction process (pose constrained-labeling + label-sensitive feature extraction), as in the model presented here.

[3] During prototyping, we have also experimented with other structured random forests including models with Potts dependencies. None of these models lead to significant improvement in labeling over the random-regression forest proposed. We have also experimented with depth features in order to better understand what is achievable with RGB features. In that case the estimates of a random forest regressor based on a depth feature similar to Kinect[16], produced comparable results with our best model on RGB features, further supporting the feasibility of our approach.

timate to rescale the image before extracting local features (rRF and lRF, if respectively only SIFT, or both SIFT and the pixel location feature is used). The scale predictor uses simple features of the foreground mask: the bounding box size (height and width), area, perimeter, eccentricity and solidity and learns to predict the distance from the camera to the pelvis bone of the body (which is available for our dataset). Images are then rescaled by that factor. The ambiguity between the left and right side (3) requires global body orientation information, for which we propose 2 solutions. The first uses confidences of a simple global direction classifier, predicting 4 body orientations and augmenting the feature vector $\boldsymbol{\varphi}$ with 4 directional confidence dimensions (rldRF). The second solution is based on a global contextual feature and appends a pose estimate at the previous labeling iteration ($16\times$ 3D joint coordinates, total of 48 dimensions) to the local features just described. This model is named rlpRF.[4]

To understand why context can be useful let us consider an example. A left-forearm of a person seen frontally may appear similar to a right-forearm if one observes the person from a short distance behind, and the context of the other body parts is not accessible (visible or used). Since the dataset is balanced with respect to the proportion of views from the front and from the back, this could create problems to the classifier. Our contextual term attempts to mitigate such issues by providing global information – the current pose estimate – to the labeling model, to further focus its predictions. While this information only seems useful if the pose estimate is accurate enough, it does work best in our experiments. Intuitively, the classifier learns not only the direction the person is facing by inspecting the positions of different joints in the image, but also learns a coarse representation of the limbs and self-occlusions.

4. Second-Order Label-Sensitive Pooling ($\boldsymbol{\Psi}$)

In this section we introduce methodology that allows us to design region descriptors that are covariant with body pose. We will rely on the intermediate 2D body part labeling (§3), aiming, at the same time, of being robust to a degree of body part mis-segmentation and noise. Each region descriptor can be obtained by appropriately dividing the region spatially into cells and pooling local descriptors, *e.g.* SIFT, over each cell independently, then concatenating the resulting vectors. Previous work has considered laying out the cells in a fixed, hand-picked configuration (*e.g.* spatial pyramids[22]), or in a fixed but learned configuration[19], both independent of the image content observed.

In contrast, we consider descriptors that provide invariance to imaging nuisance factors by pooling global statistics of the local descriptors inside multiple image regions, computed dynamically. Instead of creating a one-to-one mapping between body-parts and sets of pixels inferred to have the same label, one can view the space of possible regions to pool over as a *hierarchy*, where the coarsest level contains a region for the entire body, and the finest level has different regions for each body part. Assuming human silhouettes are available, the coarsest region is the same no matter how poorly body parts are labeled. At the other end, defining one pooling region for each body part offers potentially distinctive descriptors which may be more difficult to reliably obtain in practice. Intermediate levels of the decomposition, *e.g.* individual arms and legs, or the upper and lower body parts, offer different trade-offs between full and part visibility, discriminative power and repeatability.

In this work we propose the novel concept of *label-sensitive pooling* by defining a 'cell' (or region) configuration on the fly, based on the 2D body part labeling inferred with our learned model p. The cells are defined as potentially overlapping collections of pixels in the image, and are both *free-form* and *input-dependent*. Each region descriptor is computed by means of a *second-order label-sensitive pooling operation (O_2LP)*, by generalizing a method proposed recently in the different context of semantic segmentation[7].[5] The operator is defined as

$$\mathbf{v}(R_j, \mathbf{I}) \equiv \operatorname{vec}\Bigg(\log\Bigg(\frac{1}{|R_j|}\sum_{k\in R_j} \mathbf{s}_k \cdot \mathbf{s}_k^\top\Bigg)\Bigg) \qquad (5)$$

where $\log$ is the principal matrix logarithm, which is symmetric, hence the vectorization operator retains only the upper triangle. The dimensionality of $\mathbf{v}$ is therefore $\frac{S^2+S}{2}$. Additionally, a power transformation is applied on each individual dimension v of $\mathbf{v}$, as $\operatorname{sign}(v)\cdot|v|^h$, with $h \in [0,1]$. The final global descriptor is obtained by concatenating the descriptors $\mathbf{v}(R_j, \mathbf{I})$, for $j \in [1,\ldots,M]$, resulting in $M \times \dim(\mathbf{v})$ dimensions:

$$\boldsymbol{\Psi}(\mathbf{x}, \mathbf{I}) = [\mathbf{v}(R_1, \mathbf{I})^\top, \ldots, \mathbf{v}(R_M, \mathbf{I})^\top]^\top \qquad (6)$$

Notice the subtle dependency $\boldsymbol{\Psi}(\mathbf{x})$, as an effect of $\mathbf{v}(R_j)$ depending on $\mathbf{x}$, due to labeling $R_j = \{k|\mathbf{x}_k \in \mathcal{R}(j)\}$ (§2).

5. Human 3D Pose Estimation (f)

We investigate regression methods with simple and structured outputs for 3D human pose estimation.

[4] Notice the significant differences compared to RGB-D Kinect models: (1) we show that human body part labeling can be made feasible using RGB images only, by considering different features and spatial encodings, iteratively contextualized by 3D pose estimates (§3); (2) for RGB-D, given a body labeling and a depth map, predicting the 3D body joints is relatively easy. This is by no means the case for RGB, motivating our novel development of a label-sensitive pooling descriptor (§4).

[5] While in [7] regions were identified based on a bottom-up segmentation method, CPMC[8], here regions are defined as containing subsets of the labels, and their spatial support will vary, for different inputs, as the union of pixels with inferred labels in those subsets.

Ridge Regression (RR): In this formulation, the models for each of the D output dimensions (3D joints of the human body) will be trained independently:

$$\mathbf{f}(\mathbf{\Psi}(\mathbf{x},\mathbf{I})) = \mathbf{W}^\top \mathbf{\Psi}(\mathbf{x},\mathbf{I}) \tag{7}$$

where $\mathbf{W}$ is a matrix of size $\dim(\mathbf{\Psi}) \times 3 \times D$, with columns $\mathbf{w}_d$. Learning is performed using regularized least squares

$$\arg\min_{\mathbf{W}} \sum_{\mathbf{I}} \sum_{d=1}^{3\times D} \left\{ \|\mathbf{w}_d^\top \mathbf{\Psi}(\mathbf{x},\mathbf{I}) - \mathbf{z}_\mathbf{I}(d)\|_2^2 + \|\mathbf{w}_d\|_2^2 \right\} \tag{8}$$

where $\mathbf{z}_\mathbf{I}(d)$ is the dimension d of the ground truth pose vector $\mathbf{z}_\mathbf{I}$ in image $\mathbf{I}$, and $\mathbf{w}_d$ is the d-th row of $\mathbf{W}$. Ridge regression has closed form solution and is very efficient to train if the dimensionality of the input is not extremely high.

Kernel Dependency Estimation: For this structured predictor, we first map the targets to a RKHS, de-correlate them, then learn an input-to-target map, using ridge regression. For scalability, we use a linear kernel approximation methodology[24] to represent the non-linear lifting explicitly using a finite-dimensional approximation of size M. To learn the target map $\mathbf{\Gamma} : \mathbb{R}^{3\times D} \to \mathbb{R}^M$ we solve the following optimization problem

$$\arg\min_{\mathbf{W}} \sum_{\mathbf{I}} \sum_{d=1}^{M} \left\{ \|\mathbf{w}_d^\top \mathbf{\Psi}(\mathbf{x},\mathbf{I}) - \mathbf{\Gamma}_d(\mathbf{z}_\mathbf{I})\|_2^2 + \|\mathbf{w}_d\|_2^2 \right\} \tag{9}$$

where $\mathbf{\Gamma}_d(\mathbf{z}_\mathbf{I})$ is the d-th dimension of $\mathbf{\Gamma}(\mathbf{z}_\mathbf{I})$. Inference requires solving a pre-image problem[18]

$$\mathbf{f}(\mathbf{\Psi}(\mathbf{x},\mathbf{I})) = \arg\min_{\mathbf{z}_\mathbf{I}} \|\mathbf{W}^\top \mathbf{\Psi}(\mathbf{x},\mathbf{I}) - \mathbf{\Gamma}(\mathbf{z}_\mathbf{I})\|_2 \tag{10}$$

This can be obtained efficiently using an LBFGS optimizer, initialized using RR predictions (7). The method we employ is essentially the one in [18] except that for O_2LP we use a linear input kernel, the correct metric being already accounted for by the principal matrix logarithm (§4).

6. Experiments

Leveraging a large and diverse set of human poses within trainable systems is key to obtaining robust results, as it has become clear recently, for pose estimation systems based on RGB-D data. For our experiments we use the recently released Human3.6M (H3.6M) dataset[18], which delivers 3.6 million synchronized images and 3D body poses, as well as camera parameters and body part labeling in images. We use the set of 24 labels provided with H3.6M, out of which 20 are associated to limbs (3 for the joints of each limb and 2 for the upper and lower bones, *e.g.* the arm consists of the shoulder, elbow, wrist as well as humerus and radius), 2 for the torso (chest and abdomen), 1 for the pelvis and 1 for the head (see [18] for details). Our pose estimator uses the standard 17 joint skeleton from H3.6M. Of the entire dataset, we select a subset of 55,000 training examples and 25,000 testing examples. We refer to this training and testing subset **Human80K (H80K)**[6]. It is obtained by eliminating those 3D poses that are similar ($<$ 100 mm), in each video of H3.6M, then sampling from the remaining data uniformly. We use image data collected from all 4 cameras in H3.6M, in order to make sure that we have a sufficiently diverse set of viewing angles. We cover all the available 15 scenarios in H3.6M in roughly equal proportions of about 4,000 training examples for each motion for training and 1,800 for testing.

Our predictive model has three components: body part-labeling, label sensitive pooling and 3D human pose estimation based on the resulting descriptor. In the sequel, we will analyze each of these stages in isolation, as well as jointly.

Dense Human Body Part Labeling. As described in §3, for this task we chose a random forest (RF) which we found to outperform SVM and logistic regression classifiers, by a large margin. We trained using 150 sample patches per image, making sure that every visible label has at least one sample, then extracting the corresponding local feature. Different predictors were trained using data from the same image locations (which vary across images) to ensure they are not biased by sampling artifacts. The resulting dataset has 8 million patch descriptor examples. RF classifiers with $T = 30$ trees were trained, which took between 2.5h (our simplest RF model) to 4.5h (the most complex one), on a 8-core 2.13Ghz machine (48Gb RAM).[7]

Our results are summarized in table 1. Confusion matrices for different models are shown in fig. 2. We noticed that both introducing a pixel location and rescaling the bounding box prior to feature extraction are operations of consistent performance benefit. The same can be said about the global information: both direction and pose offer advantages, with pose providing better overall performance. Notice that pose carries a much richer context about labels than orientation disambiguation as it accounts for self-occlusion and gives a strong prior on the spatial label distribution.

Label-sensitive Pooling. Having shown that human body part labeling based on RGB images can be made effective, the next stage is the construction of descriptors $\mathbf{\Psi}$ for 3D human pose estimation, by considering a region hierarchy for label-sensitive pooling (§4). The regions were defined by considering a 4-level decomposition of the body: the first level (L1) has one region for the fully body (FB); the second level (L2) has two regions: upper (UB) and lower body (LB); the third level (L3) has 5 regions for the head

[6] Available for download and testing via the Human3.6M site.

[7] Using one compute core, in Matlab, the inference time on an image is: .47s for labeling, .81s for O_2PL feature extraction and less than .01s for pose estimation. An iteration takes less than 1.3s (we usually need 2).

	RF	rRF	lRF	rlRF	rldRF	rlpRF-RR	rlpRF-KDE
Average Accuracy (%)	58.93	60.05	63.23	63.87	69.53	72.03	73.99
Average Accuracy per Class (%)	38.84	40.08	42.70	43.55	48.92	50.40	53.10

Table 1: Comparison of classification performance (%) for random forest features. We study how augmenting our base SIFT feature further improve performance. RF is the model based on SIFT, lRF augments it with pixel location coordinates. In all cases an 'r' prefix indicate that images were scaled based on the predicted distance to the camera. The other two models are rldRF (which contains both pixel coordinates and orientation features) and rlpRF which uses the pixel location and an estimate of the 3D pose (predicted using either ridge regression (RR) or kernel dependency estimation (KDE[9, 18]) against O_2LP descriptors computed based on labels from rlRF) as an additional input feature.

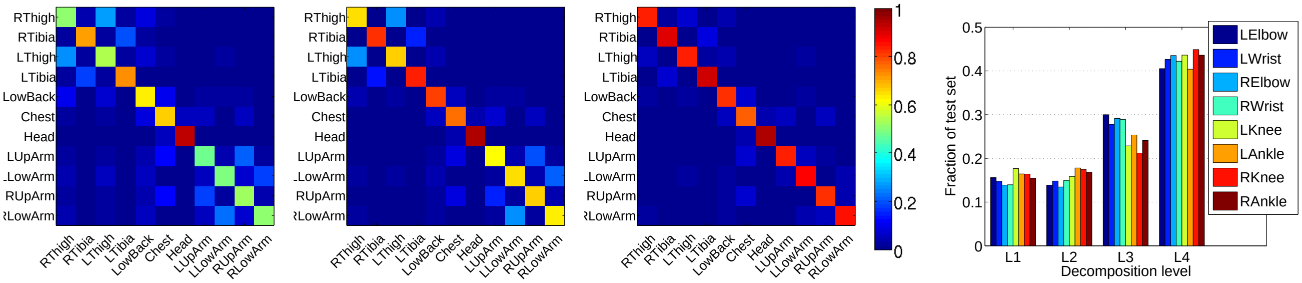

Figure 2: Confusion matrices of pixelwise 2D part labeling for different models. *First* plot shows the RF baseline, *second* the rlRF and *third* the rlpRF model. By injecting global pose information as a contextual feature in the labeling process, the rlpRF model reduces the ambiguities in assignments over the left-right side of the body, that affect the RF model. *Fourth:* Histograms showing how many times models obtained by pooling over different regions performed best at predicting different components of the 3D pose (see text).

and torso (HT) as well as each limb (LA, RA, LL, RL); the fourth level (L4) has 10 regions: torso, head and the lower and upper regions for each limb (upper arm, lower arm, thigh, calf). This results in 18 regions over which *label-sensitive pooling* will be performed. Since the O_2P descriptor is 10,000 dimensional[8] this produces a 180,000 dimensional encoding. While it is possible to learn with the high-dimensional feature vector using, *e.g*. online methods, we instead performed PCA on each region descriptor set independently, then thresholded the spectrum at the same value for all regions to obtain a 2,500 dimensional descriptor for each. This results in a 45,000 dimensional global descriptor to work with. For the KDE experiments we use a 4K dimensional approximation of a Gaussian kernel over the targets. For e^{χ_2}-HoG experiments we use a 15K dimensional approximation[24].

Automatic 3D Human Pose Prediction. Given label-sensitive descriptors $\mathbf{\Psi}$, we can obtain the regression parameters for the 3D human pose predictor in closed-form (§5). To gain insight into the effectiveness of our hierarchical representation (16 regions), we have also trained 3D human pose regressors on descriptors obtained only for those regions. Separately analyzing different decompositions offers potentially more invariant descriptors at coarser levels and more distinctive ones at finer levels. Training regressors for each region in the decomposition, $\mathcal{R}(j)$, separately, and counting how often each performs best (fig. 2, right) indicates that lowest levels are the most relevant about 50% of the time, whereas the coarsest ones are cumulatively more accurate the other 50%. While this invites other contextual output fusion methods based on model uncertainty (or *e.g.* conditional mixture of experts[31]), we still found that a combined (flat) descriptor produces accurate results overall while being fast in both training and testing.

Finally, we present the results of our automatic 3D pose estimation methods together with several baselines, in table 2 and fig. 3. Our labeling models are appended 'O_2LP' to their names in order to make clear that a label-sensitive pooling process operates over their results in order to construct the descriptors used for 3D pose estimation. The results show that given ground truth labels, all but about 10% of our errors are larger than 100mm thus validating the approach. Our automatic model, RR-O_2LP-rlpRF achieves 106mm which *represents a substantial* 33% *error reduction* over a RR-HoG baseline. This shows that improved labeling through contextual 3D pose features translates in 17% improvement over models not using it. Our rlpRF model also offers significant improvements with respect to our simpler rldRF. Note that KDE consistently improves performance.[9] Automatic labeling and 3D reconstruction

[8] While the features φ used by the random forest for labeling are obtained by appending additional information to SIFT, the label-induced region descriptor $\mathbf{\Psi}$ used for 3D human pose estimation is constructed by pooling over SIFT descriptors only.

[9] In order to compare our model with a state-of-the-art 2D pose estimation method[35] we have used the 2D ground truth joint position an-

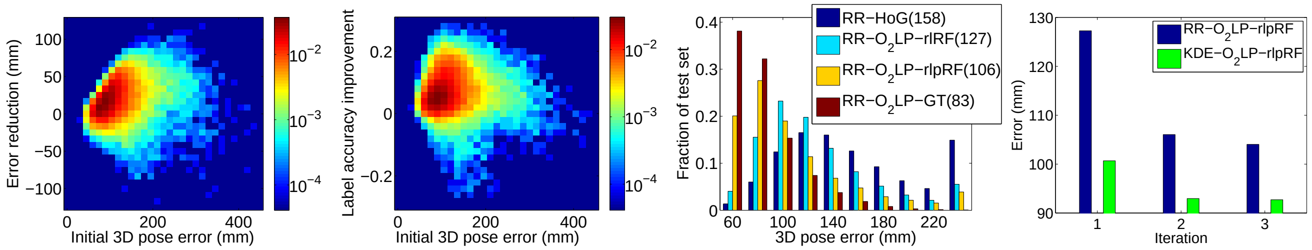

Figure 3: Left-to-right 1st and 2nd: Impact of including 3D global pose estimates in the features used for body part labeling. The first plot shows the frequency of improvement in labeling and 3D pose estimates as a function of the initial pose error. *3rd:* Error magnitude histograms showing that with perfect labels, on 90% of the test set we obtain 3D pose estimation errors less than 120mm. It also shows that our contextual labeling model enhanced with global 3D pose estimates produces $\approx$ 10% more instances where error is below 100mm. *4th:* Pose estimation performance as a function of iteration.

Method/Features	HoG	e^{χ_2}-HoG	O_2LP-GT	O_2LP-rlRF	O_2LP-rldRF	O_2LP-rlpRF
RR	158	140	83	127	115	**106**
KDE	129	112	78	100	98	**92**

Table 2: 3D pose prediction error (in mm, joint position averages) for different models. Notice the substantial improvement of our methods (fourth column based on GT labeling, last three columns automatic) over competitive regression baselines based on HOG.

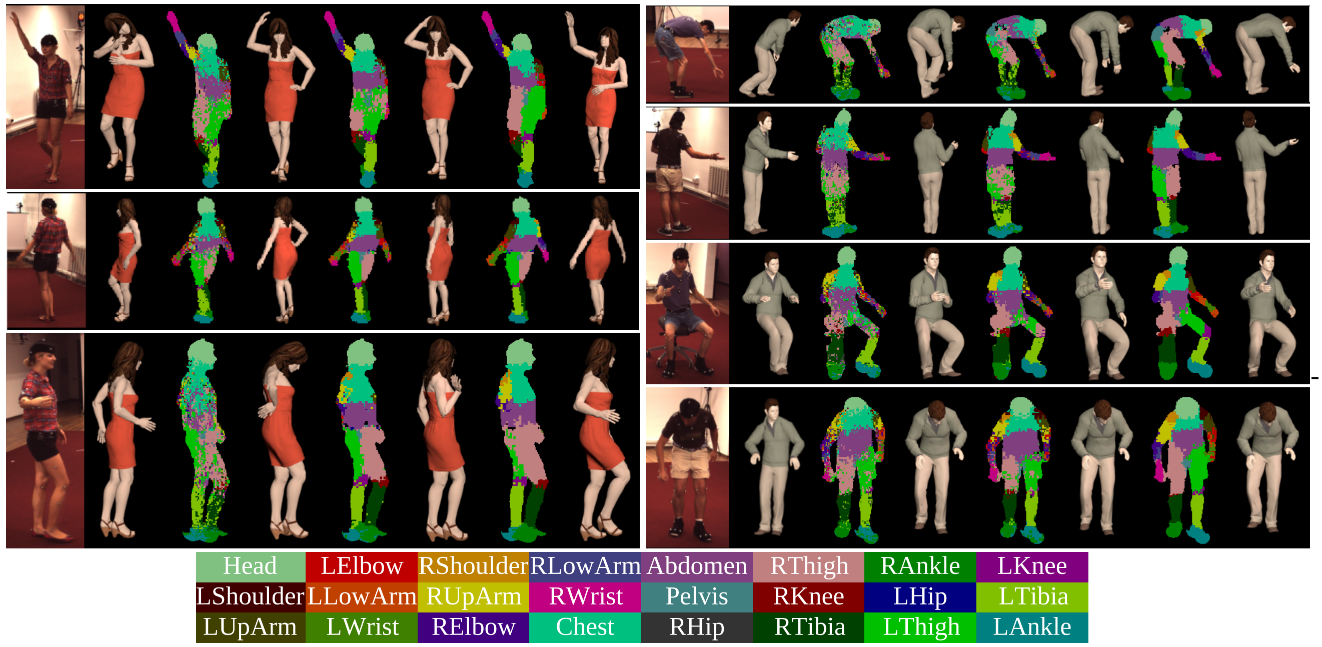

Figure 4: Examples of labeling inference and 3D human pose estimation for our different models. The columns represent from left to right: the input image, the 3D pose predicted by a HoG baseline, then labeling and estimated 3D pose for our RF, rlRF and rlpRF models.

visualizations for different models are shown in fig. 4.

7. Conclusions

We have proposed a 3D pose estimation model that decomposes into three layers: 2D human body part labeling, label-sensitive pooling over a hierarchical region decomposition of the body, and continuous-valued pose

notations from **H80K** and trained [35] on the same 55,000 images with the same foreground masks applied, as our predictors. The model had 26 parts as in the PARSE experiment[35] and otherwise all standard parameters. For this model we have obtained a PCK score 74% at (the standard, yet fairly liberal) .2 tolerance, on the test set of 25,000 images in **H80K**. For our model O$_2$P-rlpRF-KDE we have taken the 3D pose estimates and projected into the image to obtain 2D joint positions, and obtained a PCK score of 95.52% at .2 tolerance.

regression. We employ an iterative structured prediction formulation that incorporates *label-sensitive second-order pooling* over local features in order to build stable and robust pose descriptors that adapt to the human pose configuration. The diversity and annotation detail of the recently introduced Human3.6M dataset[18] makes possible to train large-scale human labeling and pose estimation methods, at large scale, for the first time. The proposed methodology operates on intensity images and leads to excellent 2D body part labeling results as well as 3D pose estimates that are significantly more accurate compared to existing competitive discriminative regression methods based on HOG. In future work, we plan to introduce additional label structure in order to handle complex backgrounds and multiple people.

Acknowledgements This work was supported in part by CNCS-UEFISCDI under CT-ERC-2012-1 and PCE-2011-3-0438. Research was performed in part while Joao Carreira was at ISR-Coimbra, supported by FCT under PTDC/EEA-CRO/122812/2010 and SFRH/BPD/84194/2012. The authors are grateful to Elisabeta Marinoiu for her generous help with the 3D human modeling and the visualizations used in the paper.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, 2004.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[3] L. Bo and C. Sminchisescu. Structured Output-Associative Regression. In *CVPR*, 2009.

[4] Y. Bo and C. Fowlkes. Shape-based pedestrian parsing. *CVPR*, 2011.

[5] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

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

[7] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

[8] J. Carreira and C. Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 2012.

[9] C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *ICML*, 2005.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013.

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

[13] S. M. A. Eslami, N. Heess, and J. Winn. The shape boltzmann machine: A strong model of object shape. In *CVPR*, 2012.

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

[15] J. Gall, B. Rosenhahn, T. Brox, and H. Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *IJCV*, 2010.

[16] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, 2011.

[17] C. Ionescu, F. Li, and C. Sminchisescu. Latent Structured Models for Human Pose Estimation. In *ICCV*, 2011.

[18] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[19] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, 2012.

[20] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007.

[21] L. Ladicky, P. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013.

[22] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[23] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 2006.

[24] F. Li, G. Lebanon, and C. Sminchisescu. Chebyshev Approximations to the Histogram χ^2 Kernel. In *CVPR*, 2012.

[25] Q. Li, J. Wang, Z. Tu, and D. P. Wipf. Fixed-point model for structured labeling. In *ICML*, 2013.

[26] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *ICCV*, 2011.

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

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

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

[30] C. Sminchisescu. *3D human Motion Reconstruction in Monocular Video. Techniques and Challenges*, volume 36 of *Human Motion Capture: Modeling, Analysis, Animation*. Springer, October 2007. ISBN 978-1-4020-6692-4.

[31] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, 2005.

[32] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *CVPR*, 2003.

[33] Z. Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008.

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

[35] Y. Yang and D. Ramanan. Articulated Human Detection with Flexible Mixtures of Parts. *PAMI*, 2013.