Depth Sweep Regression Forests for Estimating 3D Human Pose from Images

Ilya Kostrikov
ilya.kostrikov@rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Juergen Gall
gall@iai.uni-bonn.de
University of Bonn
Bonn, Germany

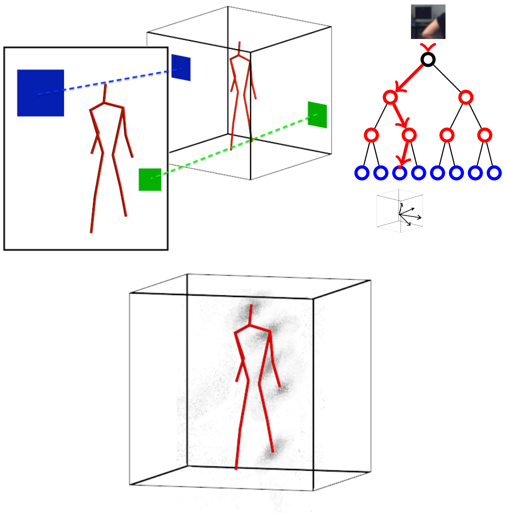

Figure 1: Illustration of a depth sweep regression forest for 3D pose estimation from a 2D image. **Top left.** Patches sampled from different depths project onto the image with different scale. **Top right.** The projected patches traverse the tree evaluating splitting functions in the intermediate nodes (black and red) until they reach a leaf node (blue). A leaf node contains 3D offsets that point to locations of a joint with associated weights. **Bottom.** Based on the offsets, the patches sampled in the 3D volume cast 3D votes for several joint locations.

Over decades estimating the human pose from still images has been an intensive research topic. In recent years, the majority of works has been focused on estimating the 2D pose, since this is already very challenging. However, many applications require the 3D pose. While some approaches estimate first the 2D pose and then reconstruct the 3D pose from the 2D pose estimate, estimating the 3D pose directly from the images is more practical since it directly solves the problem at hand. For this task, discriminative approaches that learn a mapping from image features to 3D pose, have been most successful. This is in contrast to state-of-the-art human pose estimation approaches that rely on discriminative parts and combine them within a pictorial structure model [2] that represents the human skeleton. A prominent example of these approaches is [4].

In this paper we address the problem of estimating the 3D pose from still images. However, instead of learning a regression from image features to the full pose, we regress the positions of the joints in 3D space and then infer the pose using a 3D pictorial structure framework. For regression, we rely on regression forests that have been shown to efficiently predict 2D pose from images [1]. These approaches, however, cannot be directly applied since each local image or depth feature estimates the relative positions of the joints from the feature location. While the relative position is well defined if feature and joint locations are given either in 2D or in 3D, it is not defined if the features are sampled from 2D images without depth information and the joint locations need to be predicted in a 3D world coordinate system.

Our approach consists of two parts: first, we independently estimate joint 3D location probabilities; second, we use the estimated probabilities together with the pictorial structure framework in order to infer the full skeleton. For the first part, we propose depth sweep regression forests which are regression forests that hypothesize the missing depth information of image features. For the second part, we extend the mixture of PSMs [3] for 3D inference.

In the context of pose estimation [1], a regression tree represents a mapping from the space of image patches and patch locations $\mathcal{P}\times\Omega$ to the space of probabilities over joint locations $\mathcal{X}$. In case of 2D pose estimation, we have $\Omega\subset\mathbb{R}^2$, $\mathcal{X}\subset\mathbb{R}^2$ and $\mathbf{d}(\mathbf{x},\mathbf{y})=\mathbf{x}-\mathbf{y}$. For localizing a joint j, the probabilities of all trees of a forest are averaged and summed over all patches sampled from locations $\mathbf{y}\in\Omega$:

$$\phi_j(\mathbf{x}) = \sum_{\mathbf{y}\in\Omega}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} p(j|L_T(P,\mathbf{y}))p_j(\mathbf{d}(\mathbf{x},\mathbf{y})|L_T(P,\mathbf{y})). \tag{1}$$

where $p(j|L)$ denotes the class probability of joint j stored at leaf L and $p_j(\mathbf{d}|L)$ denotes the probability of relative locations of the joint j.

In order to predict 3D joint locations from 2D images, the approach briefly described above cannot be directly applied since $\Omega\subset\mathbb{R}^2$ and $\mathcal{X}\subset\mathbb{R}^3$. The relative location $\mathbf{d}$ of a 3D joint given the 2D location of a patch, and thus (1), are not defined. We therefore propose to perform the inference in $\Omega'\subset\mathbb{R}^3$ instead:

$$\phi_j^{ds}(\mathbf{x}) = \sum_{\mathbf{y}'\in\Omega'}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} p\left(j|L_T(P,\mathbf{y}')\right)p_j\left(\mathbf{d}(\mathbf{x},\mathbf{y}')|L_T(P,\mathbf{y}')\right). \tag{2}$$

In this formulation $\mathbf{d}(\mathbf{x},\mathbf{y}')=\mathbf{x}-\mathbf{y}'$ is well defined, but the regression trees have to learn a mapping from $\mathcal{P}\times\Omega'$ to $\mathcal{X}$. This causes a problem since $\mathcal{P}\times\Omega'$ is not observed neither for training nor for testing. However, assuming that the camera projection π is known, which maps a point from Ω' to the image plane Ω, we can rephrase the problem as learning a mapping from $\mathcal{P}\times\Omega\times\mathcal{Z}$ to $\mathcal{X}$, where the appearance of a 2D patch P depends on the 2D image location and the depth z. Since we do not observe depth for training or testing, we hypothesize it by sweeping with a plane parallel to the image plane along the z-axis through a 3D volume. The patch P corresponding to the 3D point y' is then the patch centered at the projection $\pi(\mathbf{y}')\in\Omega$ and the leaf it ends depends on $z'\in\mathcal{Z}$:

$$\phi_j^{ds}(\mathbf{x}) = \sum_{\mathbf{y}'\in\Omega'}\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} p\left(j|L_T(P,\pi(\mathbf{y}'),z')\right)p_j\left(\mathbf{d}(\mathbf{x},\mathbf{y}')|L_T(P,\pi(\mathbf{y}'),z')\right). \tag{3}$$

Since the appearance of patches changes for different depth values, the maximum of (3) corresponds to a set of patches that are associated to the correct hypothesized depth values and agree on the 3D joint location.

Inferring 3D joint locations independently from 2D RGB images is prone to depth ambiguities. Many of the ambiguities, however, can be resolved by using a kinematic body model that provides information about constraints between joint locations. To this end, we use the well known pictorial structure framework [2] that provides accurate results while keeping the inference tractable:

$$p(X_J|I,\vartheta)\propto\prod_{j\in J}\left(\phi_j^{ds}(\mathbf{x}_j)\right)^{\alpha}\prod_{(i,j)\in E}\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j|\vartheta_{ij}), \tag{4}$$

As proposed in [3], we use a mixture of PS models to overcome the limitations of a single tree model. Given a set of training poses M, we cluster the relative poses by k-means and estimate the parameters of a PS model for each cluster. Inference is first performed for each PS model independently and the solution of the model with highest confidence is taken. We weight the confidence of each model by the prior probability of the model.

We compare our approach with other methods on HumanEva I and Human3.6m where our approach achieves state-of-the-art performance.

[1] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[2] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 2005.

[3] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[4] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.