Multi-task Forest for Human Pose Estimation in Depth Images

Joe Lallemand
BMW Group
München
joe.lallemand@bmw.de

Olivier Pauly
Department of Computer Science, CAMP,
Technische Universität München
pauly@cs.tum.edu

Loren Schwarz
BMW Group
München
loren.schwarz@bmw.de

David Tan
Department of Computer Science, CAMP,
Technische Universität München
tanda@in.tum.de

Slobodan Ilic
Department of Computer Science, CAMP,
Technische Universität München
Slobodan.Ilic@in.tum.de

Abstract

In this paper, we address the problem of human body pose estimation from depth data. Previous works based on random forests relied either on a classification strategy to infer the different body parts or on a regression approach to predict directly the joint positions. To permit the inference of very generic poses, those approaches did not consider additional information during the learning phase, e.g. *the performed activity. In the present work, we introduce a novel approach to integrate additional information at training time that actually improves the pose prediction during the testing. Our main contribution is a multi-task forest that aims at solving a joint regression-classification task: each foreground pixel from a depth image is associated to its relative displacements to the 3D joint positions as well as the activity class. Integrating activity information in the objective function during forest training permits a better partitioning of the 3D pose space that leads to a better modelling of the posterior. Thereby, our approach provides an improved pose prediction, and as a by-product, can give an estimate of the performed activity. We performed experiments on a dataset performed by 10 people associated with the ground truth body poses from a motion capture system. To demonstrate the benefits of our approach, poses are divided into 10 different activities for the training phase. Results on this dataset show that our multi-task forest provides improved human pose estimation compared to a pure regression forest approach.*

1. Introduction

Estimation of the 3D generic human body pose is a widely studied problem in the Computer Vision community. Earlier works attempt to solve it using monocular camera images [1, 12, 25]. However, it is known that, due to the projection ambiguity, this is a heavily under-constrained problem. For this reason, body pose estimation has also been constrained by particular activities that have been learned beforehand. However, this was mainly used for generative articulated human body tracking [5, 21, 26, 27] and not for the discriminative estimation of the 3D body pose independently in every frame.

Recently, with the advent of depth sensors like the Kinect and ToF cameras, novel approaches to human body tracking and pose estimation from depth data have been introduced. These approaches are classified as generative [2, 4, 16, 18], discriminative [8, 10, 17, 22] or a combination of both [9, 19, 23]. Lately, researchers from Microsoft proposed an approach based on decision forests [22] that classifies depth pixels belonging to the different body parts. As the forest output does not provide any global information on the body pose, this approach suffers from limitations in the case of occlusions. Girshick *et al*. [10] extended this approach and used regression forests to learn the function that maps each depth pixel to its offsets from all body joints. This allowed a better prediction of the joint location and handling of self-occlusions. While these methods aim to estimate the generic 3D human body pose, the approach of [21] simultaneously returns the body pose and the activity class. This method combines manifold learning of human activities and particle filtering for human body tracking and activity recognition.

In this paper, we address a similar objective but in a discriminative setting where we want to improve the estimation of the 3D body pose by integrating activity information at training time. Inspired by the human body pose estimation approach based on regression forests [10] and the manifold learning approach of [21], we propose to jointly learn generic human body poses and human activity in multi-task

forests by associating each foreground pixel in a depth image with its offsets from all body joints as well as the current activity class. The main contribution is the integration of the activity information available at training time by formulating the problem as a joint regression-classification task, *i.e*. regression of the 3D human body pose and classification of the performed activity. Similar to [11], we propose a multi-task forest to tackle jointly this regression-classification task. As we would like to infer any type of poses, including those that do not belong to the set of activities seen during training, we assume independence between joint offsets and activities. While it seems that this makes the search space even larger, the integration of the activity information encourages a faster clustering of training samples into leaves that are consistent in terms of both poses and activities. This allows to model both pose and activity probability distributions in the leaves and to employ different prediction strategies. This paper addresses the following scenarios: (i) predicting the pose while ignoring the activity and (ii) predicting jointly both the pose and the activity.

Our approach is evaluated on a dataset containing 10 people performing 10 activities such as golf, football or boxing. The results show that the multi-task forests perceptibly improve the body pose estimation for test sequences that exhibit motions from activities used during the pose learning phase.

1.1. Related Work

Recovering 3D human body pose from images and videos has been extensively studied [5, 14]. However, many interesting developments have emerged in the last years and dramatically advanced the field of human motion capture. Here, we concentrate only on works that consider monocular color or depth cameras. Using color monocular cameras is particularly challenging because of the depth ambiguities. Many approaches that address this problem [1, 12, 25] are discriminative and learn a mapping from visual observations to articulated body configurations. Agarwal and Triggs [1] use shape context descriptors to describe 2D silhouettes and Relevance Vector Machine regression to learn the mapping from silhouettes to 3D human poses. Urtasun and Darrell [25] use Gaussian Process (GP) models which offer a general framework for probabilistic regression and have been shown to generalize well for small training data sets. Kanaujia et al. [12] employ semi-supervised learning of hierarchical image descriptions in order to better tolerate deformation, misalignment and clutter in the training and test sets.

Approaches for 3D human body pose estimation based on discriminative models have inspired recent methods that use depth sensors like Kinect and Time-of-Flight (ToF) cameras [9, 10, 17, 22]. Plagemann *et al*. [17] propose an approach for body part detection in ToF camera data based on interest points computed using geodesic extrema and boosted classifiers. Ganapathi *et al*. [9] rely on this body part detection approach and couple it with a generative skeleton-based method for real-time human body tracking. Shotton *et al*. [22] use decision forests to learn the mapping from depth pixel locations to body part labels. However, their approach does not provide 3D body poses directly, which have to be inferred from the classification output. Errors in the classification caused by self-occlusions result in wrongly estimated poses. Inspired by Implicit Shape Models of [3] and [8, 15] that learn offsets from every pixel in the image to the locations of body joints, Girshick et al. [10] use regression forests to directly estimate 3D positions of the body joints from Kinect depth images. Their approach is more robust than the previous body part classification strategy [22]. In [24], Taylor *et al*. introduce a new approach using regression forests to map each pixel to its most likely correspondance on a human body manifold. Afterwards, an energy function is optimized on this model in order to best fit the observations. In the context of hand pose estimation, Keskin *et al*. proposed in [13] a two layers strategy where they first cluster their observations based on the 2D hand shape and then train an "expert" regressor for pose estimation within each cluster.

Unlike the methods discussed above that estimate generic 3D human body poses from images without having information about prior activities, there are many approaches that strongly rely on activity priors in order to track human poses in images [5, 26, 27] and depth data [7, 20]. All propose the integration of an activity prior in order to improve human body tracking. The methods rely on manifold learning in order to learn particular activities and then exploit the resulting low-dimensional motion models to constrain the human body pose estimation problem. However, since these are generative tracking methods, they do not generalize well to poses that lie far from the manifold and body poses are not estimated directly from every single image independently.

Up to our knowledge, human pose estimation has not yet been addressed with a multi-task forest that integrates regression from individual depth image pixels to body joint locations simultaneously with activity classification. In the context of multiple objects segmentation [11], Glocker *et al*. integrated continuous spatial information in addition to the discrete object class labels within a joint regression-classification forest model. By applying this model to multiple organs segmentation, they could demonstrate improved class predictions. We propose to tackle the pose estimation problem in a similar way by learning both activity priors and a probabilistic mapping from depth pixels to joint locations.

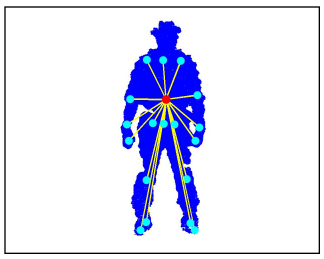


Figure 1. The points of the foreground mask $\mathcal{M}$ are represented in blue and the joint locations $\mathbf{j}_i$ in cyan. Offsets ψ_i from a pixel (in red) to all joint locations are depicted by yellow vectors.

2. Method

This section presents the novel multi-task forest model which permits to efficiently integrate additional information in the form of an activity prior to improve 3D body pose estimation. Taking advantage from the mixed output provided by the model, different pose estimation strategies are introduced depending on whether or not activity information should be considered at test time.

2.1. Problem Statement

For a 2D depth image, let $\mathbf{D}$ be an intensity function $\mathbf{D} : \Omega \rightarrow \mathbb{R}$ where $\Omega \subset \mathbb{R}^2$ is the image domain. The actor, whose pose should be estimated is given by a foreground region $\mathcal{M} \subset \Omega$. A 3D body pose is defined as a set of 3D locations $\mathbf{j}_i \in \mathbb{R}^3$ for N body joints, written as a vector $\mathbf{J} = \left[\mathbf{j}_1^\top, \cdots, \mathbf{j}_N^\top\right]^\top$. Let $\mathcal{J} \subset \mathbb{R}^{3N}$ denote the space of all possible 3D body poses. For a given depth image $\mathbf{D}$, the goal is to estimate the 3D pose $\mathbf{J} \in \mathcal{J}$.

Following a probabilistic regression approach, each pixel $\mathbf{x} = (x, y)^\top \in \mathcal{M}$ predicts the *relative offsets* $\mathbf{\Psi} = \left[\psi_1^\top, \cdots, \psi_N^\top\right]^\top$ pointing from its corresponding 3D location $\mathbf{X}$ of the pixel to the 3D location of all joints in the body pose $\mathbf{J}$. Each entry of the offsets vector $\mathbf{\Psi}$ is defined as $\psi_i = (\mathbf{j}_i - \mathbf{X})$ as illustrated by Fig. 1. The contribution of each pixel $\mathbf{x}$ to the full body pose prediction is estimated using maximum a posteriori:

$$\hat{\mathbf{\Psi}} = \underset{\mathbf{\Psi} \in \mathbf{\Lambda}}{\operatorname{argmax}} P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D}). \tag{1}$$

The space $\mathbf{\Lambda} \subset \mathbb{R}^{3N}$ spanned by all offsets to 3D joint locations is very large. Thus, the search for a good estimate of the 3D pose is challenging. While "common" poses may be well predicted, poses that are under-represented with respect to a training set may be "averaged out" in the modeling of such a posterior distribution. Nevertheless, these less common poses are often characteristic for a certain type of activity. Thus, integrating activity information permits to better model the posterior, and thereby can provide better pose prediction.

Considering a set of K activities $\mathcal{A} = \{\mathbf{a}_k\}_{k=1}^K$, our main contribution is to integrate activity information by reformulating the stated problem as the joint estimation of pose and activity:

$$(\hat{\mathbf{\Psi}}, \hat{\mathbf{a}}) = \underset{\mathbf{\Psi} \in \mathbf{\Lambda}, \mathbf{a} \in \mathcal{A}}{\operatorname{argmax}} P(\mathbf{\Psi}, \mathbf{a}|\mathbf{x}, \mathbf{D}). \tag{2}$$

In contrast to [11], we assume conditional independence between the continuous and discrete variables, *i.e.* joint offsets and activities, to permit inference of generic poses. We can thus rewrite this joint distribution as:

$$P(\mathbf{\Psi}, \mathbf{a}|\mathbf{x}, \mathbf{D}) = P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D}) P(\mathbf{a}|\mathbf{x}, \mathbf{D}). \tag{3}$$

To approximate this distribution, we propose a multi-task forest framework which integrates both $P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})$ and $P(\mathbf{a}|\mathbf{x}, \mathbf{D})$ within the same model.

The first step is to introduce a new objective function within the tree training which integrates both pose and activity information. The second step is to model both terms $P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})$ and $P(\mathbf{a}|\mathbf{x}, \mathbf{D})$ in the leaves which permits the computation of the joint probability distributions using Eq. 3. The versatility of this model is the adaption of the previously mentioned different prediction scenarios.

In the following sections, we shortly present the feature descriptor which characterizes the context of a pixel $\mathbf{x}$ given the depth image $\mathbf{D}$ as earlier proposed in [22]. We then describe the training and testing phases of the multi-task forest model.

2.2. Feature Descriptor

Our approach adopts the feature descriptor proposed by Shotton *et al*. [22], each dimension consisting of a simple depth comparison at predefined offsets from the considered pixel. Before feature extraction, we identify the image region $\mathcal{M} \subset \Omega$ that corresponds to the person in the foreground. For this purpose, we subtract a previously acquired static background image $\tilde{\mathbf{D}}$ of the scene from the depth image $\mathbf{D}$. For a given pixel $\mathbf{x} \in \mathcal{M}$ and offsets $\theta = (\mathbf{u}, \mathbf{v})$, the depth feature $\mathbf{f} \in \mathcal{F}$ is defined as

$$f(\mathbf{x}, \theta) = \mathbf{D}\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{D}(\mathbf{x})}\right) - \mathbf{D}\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{D}(\mathbf{x})}\right) \tag{4}$$

where $\mathcal{F}$ denotes the space spanned by these features and $\mathbf{D}(\mathbf{x})$ returns the depth information at pixel $\mathbf{x}$. The offsets are divided by the depth at the given pixel to achieve depth-invariance. Moreover, it measures the depth at the same distance in *world coordinates* regardless of the distance between the person and the camera.

2.3. Multi-task Forest

In this section, we describe our novel forest model that aims at learning the joint probability distribution $P(\mathbf{\Psi}, \mathbf{a}|\mathbf{x}, \mathbf{D})$. With an ensemble of decorrelated trees, this model permits to: (1) efficiently partition the feature space $\mathcal{F}$ described in the previous section and (2) estimate this

joint distribution in each part of this space. Since $\mathbf{\Psi} \in \mathbf{\Lambda}$ and $\mathbf{a} \in \mathcal{A}$ are a multivariate continuous and a categorical random variable, respectively, the proposed forest model relies on a hybrid regression-classification strategy. To this end, we use an objective function which aims to create leaves that are consistent not only in terms of joint offsets, but also in terms of activity. Let $\mathcal{T} = \{\mathbf{T}_t\}_{t=1}^{T}$ denote a multi-task forest consisting of T trees $\mathbf{T}_t$. In the next section, we describe how each tree is trained and how the pose and activity estimation is inferred from the trees.

2.3.1 Forest Training

Let us consider a training set built from a bootstrap of pixels extracted in different depth images, described by their feature vectors and associated to their joint offsets and activity labels. Each training instance is given by $\mathcal{X}^{(i)} = (\mathbf{f}^{(i)}, \mathbf{\Psi}^{(i)}, \mathbf{a}^{(i)})$, for $i \in \{1, \ldots, N_{\text{train}}\}$, where N_{train} is the number of training samples. Note that $N_{\text{train}} = N_{\text{pix}} \cdot N_{\text{images}}$. N_{images} is the number of images and N_{pix} is the number of pixels bootstrapped from the foreground in each image. In the forest model, each tree $\mathbf{T}_t$ aims to create an independent partition of the high-dimensional feature space $\mathcal{F}$. A tree is defined as a directed acyclic graph such that each node consists of a decision function $g_{\mathbf{v},\tau}$:

$$g_{\mathbf{v},\tau} = (\mathbf{f} \cdot \mathbf{v} \geq \tau) \tag{5}$$

where $\mathbf{v}$ is a vector of dimensionality $\dim(\mathcal{F})$ and $\tau \in \mathbb{R}$ is a threshold. Note that $\mathbf{v}$ has a single non-zero entry, *i.e.* $||\mathbf{v}||_0 = 1$. Depending on the result of this decision function, an incoming pixel $\mathbf{x}$ described by $\mathbf{f}$ is sent downward the tree to the left or the right child node. Here, the role of the vector $\mathbf{v}$ is to select a single feature dimension to perform the decision, thus yielding axis-aligned splits in $\mathcal{F}$.

Let Δ be the set of training instances reaching the current node, and let Δ_l and Δ_r be the subsets sent to the left and right child nodes. The choice of $(\mathbf{v}, \tau)$ is optimized following a greedy strategy: a set Γ of candidates is generated randomly and the best $(\mathbf{v}^*, \tau^*)$ are selected by maximizing the information gain:

$$(\mathbf{v}^*, \tau^*) = \underset{(\mathbf{v},\tau)\in\Gamma}{\operatorname{argmax}} \left(\mathbf{H}(\Delta) - w_l \mathbf{H}(\Delta_l) - w_r \mathbf{H}(\Delta_r)\right), \tag{6}$$

with $w_l = |\Delta_l|/|\Delta|$ and $w_r = |\Delta_r|/|\Delta|$. Based on Shannon's entropy, we define $\mathbf{H}$ as

$$\begin{aligned} \mathbf{H} = &-\alpha \cdot \int_{\mathbf{\Lambda}} P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D}) \log\left(P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})\right) d\mathbf{\Psi} \\ &-(1-\alpha) \cdot \sum_{\mathcal{A}} P(\mathbf{a}|\mathbf{x}, \mathbf{D}) \log\left(P(\mathbf{a}|\mathbf{x}, \mathbf{D})\right), \end{aligned} \tag{7}$$

$\alpha \in \mathbb{R}$ being a weight that controls the trade-off between the first term, which is the *body pose regression* objective, and the second term, which is the *activity classification* objective. Unlike [11] where the authors assume a fixed dependence between both terms, we allow independence between both terms and allow both to give different contributions to the information gain, modelled by the weight α. $P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})$ is modelled by a multivariate Gaussian distribution with its mean $\mu^{(\Delta)}$ and covariance $\mathbf{\Sigma}^{(\Delta)}$ estimated from the subset Δ:

$$P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D}) = \mathcal{N}(\mathbf{\Psi}|\mu^{(\Delta)}, \mathbf{\Sigma}^{(\Delta)}). \tag{8}$$

The body pose regression objective can then be simplified to $\frac{1}{2}\log\left((2\pi e)^{3N}|\mathbf{\Sigma}^{(\Delta)}|\right)$ after solving the integral. Note that we do not keep the full covariance matrix of Eq. 8 but only the diagonal elements to enforce independence between all the joints in a body pose which is similar to the simplification that has been applied in [6, 10]. This ensures that the positions of different limbs are independent *e.g.* that an arm is independent of the position of a leg. The activity posterior $P(\mathbf{a}|\mathbf{x}, \mathbf{D})$ is modeled using a histogram where each entry is estimated as the normalized count of training instances belonging to a given activity:

$$P(\mathbf{a}|\mathbf{x}, \mathbf{D}) = \frac{|\left\{\mathcal{X}^{(i)} \in \Delta,\ \mathbf{a}^{(i)} = \mathbf{a}\right\}|}{|\Delta|} \tag{9}$$

By iteratively splitting the nodes, a tree is grown until: (i) a maximal depth has been reached; (ii) the number of training instances falls below a predefined threshold; or, (iii) the information gain is equal to zero, *i.e.* no good split candidate has been found. The training of each tree $\mathbf{T}_t$ finally results in a set of leaves which defines the partition over the feature space $\mathcal{F}$. Intuitively, this hybrid objective function encourages the creation of clusters in the leaves that are consistent in terms of both joint offsets and activity.

Now, in each leaf, both distributions $P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})$ and $P(\mathbf{a}|\mathbf{x}, \mathbf{D})$ are estimated. First $P(\mathbf{\Psi}|\mathbf{x}, \mathbf{D})$ uses the mean-shift on the set of points reaching it and retains the points that contribute to the main mode. The main mode and the corresponding weight, which is equal to the number of points voting for it, as well as the histogram modelling $P(\mathbf{a}|\mathbf{x}, \mathbf{D})$ are finally stored in the leaf.

2.3.2 Forest Prediction

Let $\mathbf{D}$ be a previously unseen depth image and let $\mathcal{M} = \left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N_{\text{test}}}$ be the set of N_{test} pixels belonging to the foreground region. For each of these pixels, we extract the corresponding feature vectors and push them through each of the T trees of the forest. Once a pixel $\mathbf{x}^{(i)}$ reaches a leaf in tree $\mathbf{T}_t$, we gather the stored joint offset distribution as well as the activity posterior and confidence per joint. We use the stored main mode in the leaf as an estimate for the joint offset vector $\hat{\mathbf{\Psi}}_t^{(i)}$ for the current pixel,

$\mathbf{c}_t^{(i)} = [\mathbf{c}_{t,1}^{(i)} \cdots \mathbf{c}_{t,K}^{(i)}]$ denotes the confidence for each joint and $\hat{\mathbf{a}}_t^{(i)}$ denotes the most probable activity within this leaf. For better readability, we denote $\mathbf{M} = \{\hat{\mathbf{\Psi}}_t^{(i)}\}$, $\mathbf{C} = \{\mathbf{c}_t^{(i)}\}$ and $\mathbf{A} = \{\hat{\mathbf{a}}_t^{(i)}\}$ as the joint offsets, confidence and activity posterior contributions from all pixels and trees, respectively, where $i = 1, \ldots, N_{\text{test}}$ and $t = 1, \ldots, T$. In the next section, we describe how to aggregate these contributions over all pixels to estimate the body pose.

2.3.3 Pose and Activity Estimation

As our forest model provides rich outputs for each pixel, we can think of two different aggregation scenarios: (1) predicting the pose ignoring the activity; and, (2) predicting jointly the pose and the activity. We are given a set of pixel predictions for joint offsets $(\mathbf{M}, \mathbf{C})$ and activities $\mathbf{A}$. The strategy is to select the best subset of pixel/tree contributions, by choosing the 10% most confident contributions given by $\mathbf{C}$ whose offsets from the pixel are within a predefined distance threshold λ. Now, let us briefly describe our two different scenarios:

- **Pose estimation ignoring activity posteriors:** In this case, activity information is only used during the training phase to provide a better clustering. During the test phase, we only consider the joint offset predictions and their respective confidence $(\mathbf{M}, \mathbf{C})$. The position of each joint $\mathbf{j}_k$ is then inferred by aggregating the most confident contributions, i.e. by calculating the main mode using meanshift to further increase robustness to outliers:

$$\mathbf{j}_k = \text{meanshift}(\{\hat{\psi}_{t,k}^{(i)} + \mathbf{X}^{(i)} \mid \hat{\mathbf{\Psi}}_t^{(i)} \in \mathbf{M}\,, \; \|\hat{\psi}_{t,k}^{(i)}\| \leq \lambda,\ \mathbf{c}_{t,k}^{(i)} \geq \gamma\}), \tag{10}$$

  where $\hat{\psi}_{t,k}^{(i)}$ contains the 3D components from the joint offset vector $\hat{\mathbf{\Psi}}_t^{(i)}$ that correspond to the k-th joint. $\gamma \in \mathbb{R}$ is a threshold for selecting the most confident predictions, $\mathbf{c}_{t,k}^{(i)}$ being the components of $\mathbf{c}_t^{(i)}$ for the k-th joint.

- **Joint pose and activity estimation:** Considering that a person may perform one of several activities which were learned during training, we can use the activity prior to enhance the pose estimation. First, the current activity is inferred by aggregating all activity estimates $\mathbf{A}$ and choosing the dominant class which we will denote $\mathbf{a}$. Once the activity is identified, all pose contributions from differing activities are eliminated and the remaining predictions are processed as follows:

$$\mathbf{j}_k = \text{meanshift}(\{\hat{\psi}_{t,k}^{(i)} + \mathbf{X}^{(i)} \mid \hat{\mathbf{\Psi}}_t^{(i)} \in \mathbf{M}, \; \|\hat{\psi}_{t,k}^{(i)}\| \leq \lambda,\ \hat{\mathbf{a}}_t^{(i)} = \mathbf{a}, \mathbf{c}_{t,k}^{(i)} \geq \gamma\}). \tag{11}$$

3. Experiments and Results

In the following section, we demonstrate experimentally the benefits of integrating the activity prior into the training phase of our forest to improve the body pose estimation. We start by describing the datasets we recorded and continue by providing quantitative results for the prediction scenarios outlined above. We also compare our results to our own implementation of the regression forest described in [10]. We opted for the regression strategy instead of the classification strategy for training the tree structure, because creating the ground truth needed for the classification is not within the scope of this work.

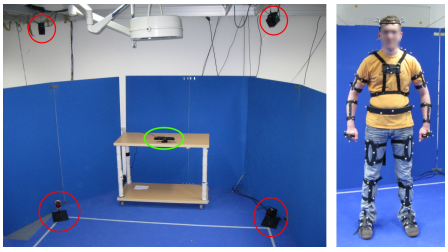


Figure 2. On the left: Setup used for the data acquisition. The red circles show 4 of the 8 infrared cameras from the motion capture system, and the green circle the position of the Kinect. On the right: One testing candidate wearing all 15 markers provided by the motion capture system.

3.1. Data Acquisition

We recorded a dataset consisting of Kinect depth images at a resolution of 640×480 pixels, synchronized with a marker-based optical motion capture system[1]. As seen in Fig.2 on the left, our system consists of 8 infrared cameras located around the recording area and provides the 3D locations of the markers shown in Fig.2 on the right with an average error of less than 0.5mm. After a preliminary body calibration for each person, the 3D locations of $N = 18$ body joints is computed in relation to the recorded marker positions for each frame. These recordings are used for both training and ground truth in our evaluations. For a total of 10 persons, we recorded three sequences for each person containing $K = 10$ activities: golfing, kicking, boxing, bowling, archery, kneeling, tennis, touching your head and moving a horizontal and vertical slider, as in a virtual user interface. One sequence consists of approximately 2,500 frames, captured at 30 frames per second. An example for each activity is given in Fig. 3 together with the estimated body pose.

[1] http://www.ar-tracking.com/

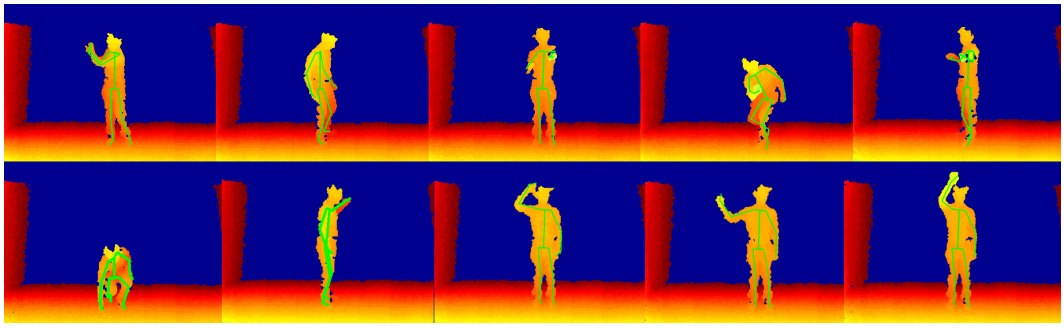


Figure 3. Depth data of a person performing all 10 activities in the following order: golfing, kicking, boxing, bowling, archery, kneeling, tennis, touching your head and moving a horizontal and vertical slider. The estimated body pose is shown in green.

3.2. Pose Estimation Accuracy

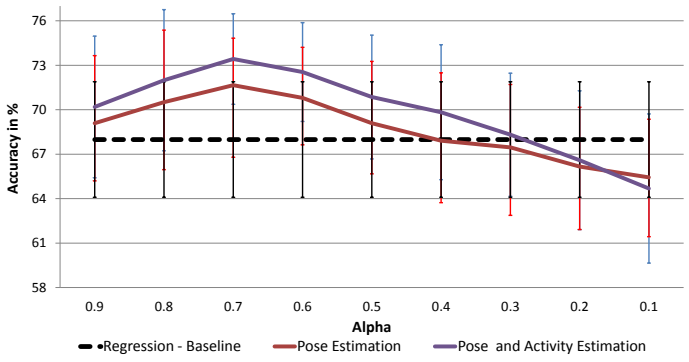
To evaluate the pose estimation capabilities of our proposed algorithm, we relied on two measures. First, the *distance error* e_{dist} represents the average metric deviation of the predicted joint positions from the ground truth, either per joint or averaged over all joints. Second, the *accuracy* e_{acc} represents the percentage of frames in the testing sequences where the predicted body pose accurately matches the ground truth. In this context, we consider a pose to be predicted accurately if all joint locations in this pose deviate less than 0.1 meters from the corresponding ground truth locations.

All results presented in the remainder of this paper were obtained in a series of leave-one-person-out validation experiments. Each of the 10 recorded people was consecutively omitted during training and used to test the resulting forest. The number of trees have been set to 3, their maximum depth to 20, and a bandwidth of 0.06m was chosen for the meanshift. Initially, we investigate the influence of the weighting parameter α that balances contributions of the pose and activities when training the forest (see Eq. 7). Fig. 4 shows the mean accuracy for each considered value of α, averaged over all 10 testing sets. Note that the same multi-task forest was trained using the activity prior. The difference is only in how the tree outputs are combined in the testing phase. Values are provided for the two voting scenarios employed in the testing phase: (i) predicting the pose ignoring the activity and (ii) predicting jointly both the pose and the activity. The best results were achieved for $\alpha = 0.7$ that shows an increase of 3.78% in precision compared to the approach of Girshick *et al*. [10] which in that case is denoted by $\alpha = 1$, as no activity classification is performed, and which is represented by the black dashed line in Fig. 4. In case of simultaneous pose and activity estimation, the accuracy increases by 5.78%. Note that for values of α starting at 0.4, the accuracy starts decreasing even below baseline regression approach. Manual inspection of the trees has shown, that from a given weighting in favor of the activity classification, the objective function starts performing the classification and the regression on the pose in a serial manner, instead of simultaneously. Consequently, the maximum depth criterion is reached in parts of the trees during the training before the poses are properly clustered in the leaves. This results in a decrease of overall accuracy.

In Fig. 5, the distance error e_{dist} is measured per joint and the comparison of the multi-task forest trained using the best $\alpha = 0.7$ with the results of the pure regression on the pose is shown. Again, all results were averaged over all different testing candidates. As anticipated after the accuracy analysis in Fig. 4 an increased precision has been obtained for the majority of the joints. The largest errors are observed in the arms, which have the highest variability and thus are most difficult to predict correctly. Fig.4 shows the confusion matrix for the activity outputs provided by the forest. While activities such as golf, football, bowling or kneeling are in general well recognized, others such as boxing, archery, touching your head or sliders are more difficult to estimate, and this because of several reasons. First, these activities involve short movements, which results in less frames within the training set; second, they contain similar poses which leads to a higher confusion between those activity classes.

4. Discussion

In our multi-task forest model, we chose to assume independence between body pose and activity class. While one could argue that this assumption might be too simplistic, this permits us to end up with a single 3D joint offsets and a class posterior distribution in each leaf. Modeling dependence between both would require a 3D joint offsets distribution **per class** which would have several disadvantages. First, the classification term in the objective function encourages the separation between the classes. This makes the computation of class-specific covariance matrices numerically unstable when only few training instances from a given class are available. Moreover, marginalization over



	Golf	Football	Boxing	Bowling	Archery	Kneeling	Tennis	Touch Head	Horiz. Slider	Vertical Slider
Golf	85,18	0,25	0,66	0,98	0,00	4,18	1,88	0,00	5,16	1,72
Football	0,43	77,52	1,43	4,62	0,00	6,75	2,50	0,00	4,20	2,55
Boxing	2,17	8,67	34,62	0,00	3,41	0,11	9,53	0,11	14,90	26,49
Bowling	1,48	4,52	0,41	57,88	6,16	20,53	2,96	4,60	1,48	0,00
Archery	3,19	7,97	5,23	6,29	55,98	0,00	6,02	8,59	3,45	3,28
Kneeling	2,01	12,21	0,00	3,34	0,00	80,43	0,00	1,84	0,17	0,00
Tennis	7,46	14,24	2,18	4,98	0,12	0,00	60,88	0,62	8,02	1,49
Touch Head	1,24	7,95	6,63	0,33	0,00	0,00	0,00	37,12	40,43	6,30
Horiz. Slider	6,64	4,82	4,82	1,91	0,00	0,00	1,09	1,18	66,33	13,19
Vertical Slider	3,56	3,67	1,15	4,25	0,00	1,38	5,51	3,67	7,58	69,23

Figure 4. On the left: Mean accuracy for all alphas averaged over all 10 testing candidates, the dotted black line shows the baseline regression forest, the red line shows the simple pose estimation voting scheme, the blue line shows the simultaneous activity and pose estimation voting scheme. On the right: Activity confusion matrix for alpha = 0.7 averaged over all testing persons

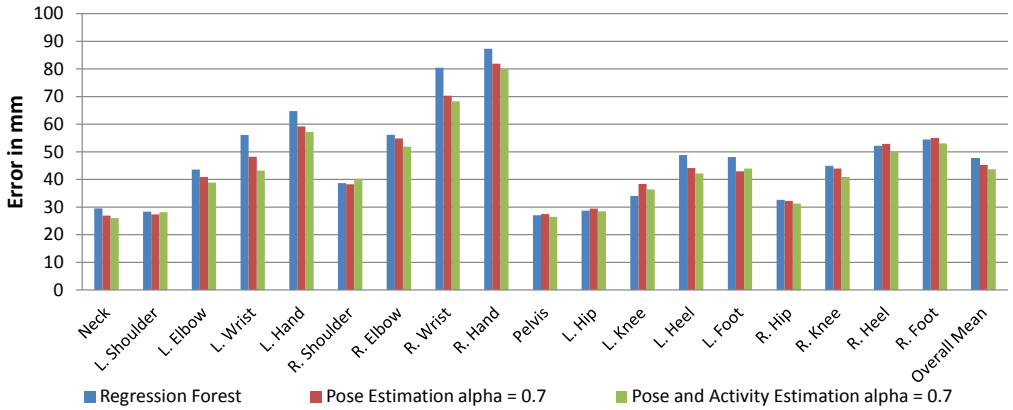


Figure 5. Comparison of the mean distance error e_{dist} per joint averaged over all 10 testing candidates. The blue bars shows the error for the pure regression on the human poses approach, and the second set of bars indicated in red show the results our our multi-task forest trained with the activity priors and for the best choice of $\alpha = 0.7$.

the classes is necessary to predict generic poses encountered in unknown activities. In contrast, as in our first scenario, one can ignore the activity information stored in the leaf and still perform generic 3D pose estimation. Of course, we do not argue that our approach can estimate poses that have not been seen during training, but we would like to emphasize on the importance of performing a good pose clustering during the training phase. In these experiments, results suggest that integrating additional activity information during the forest training permits to improve the body pose clustering and thereby, the quality of the predictions.

5. Conclusion and Future Work

In this paper, we introduced a novel discriminative approach integrating activity information at training time to improve 3D human pose estimation. To this end, we proposed to formulate the problem as a regression-classification task, in which each pixel of a depth image is associated not only to its offsets to all 3D joint positions but also to a class of activity. Therefore, we used a multi-task forest which optimizes a mixed classification-regression objective function during training. This enabled us to use two new prediction scenarios, one of which can effectively ameliorate the human body pose estimation when the activity is a priori learned. As a side effect, an estimation of the currently performed activity is given on a per frame base. Using a motion capture system, we created a dataset including 10 activities performed by 10 people with its corresponding ground truth 3D body poses. In our experiments, we could show the benefits of our approach that permits to improve human pose estimation compared to a pure regression forest approach. In future work, we want to analyze the effects of a larger training dataset on the results of the body pose estimation in terms of integrating more people into the dataset and by introducing new, more diverse activities.

References

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

and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–882. IEEE, 2004. 1, 2

[2] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *IEEE 13th International Conference on Computer Vision (ICCV)*, pages 1092–1099. IEEE, Nov. 2011. 1

[3] B.Leibe, A. Leonardis, and B. Schiele. An implicit shape model for combined object categorization and segmentation. In *Toward Category-Level Object Recognition*, pages 508–524, 2006. 2

[4] A. Bleiweiss and E. E. G. Kutliroff. Markerless motion capture using a single depth sensor. In *ACM SIGGRAPH ASIA 2009 Sketches*, SIGGRAPH ASIA '09, pages 20:1–20:1, New York, NY, USA, 2009. ACM. 1

[5] A. Elgammal and Lee. The role of manifold learning in human motion analysis. *Computer*, 36:25, 2008. 1, 2

[6] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1022–1029. IEEE, 2009. 4

[7] J. Gall, A. Yao, and L. J. V. Gool. 2d action recognition serves 3d human pose estimation. In *ECCV (3)*, pages 425–438, 2010. 2

[8] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2188–2202, 2011. 1, 2

[9] V. Ganapathi, C. Plagemann, S. Thrun, and D. Koller. Real time motion capture using a single time-of-flight camera. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, June 2010. 1, 2

[10] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, 2011. 1, 2, 4, 5, 6

[11] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi. Joint classification-regression forests for spatially structured multi-object segmentation. In *IEEE European Conference on Computer Vision (ECCV)*, 2012. 2, 3, 4

[12] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007. 1, 2

[13] C. Keskin, F. Kıraç, Y. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. *Computer Vision–ECCV 2012*, pages 852–863, 2012. 2

[14] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006. 2

[15] J. Müller and M. Arens. Human pose estimation with implicit shape models. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, ARTEMIS '10, pages 9–14, New York, NY, USA, 2010. ACM. 2

[16] Y. Pekelny and C. Gotsman. Articulated object reconstruction and markerless motion capture from depth video. *Comput. Graph. Forum*, 27(2):399–408, 2008. 1

[17] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, Anchorage, Alaska, USA, 2010. 1, 2

[18] J. Rodgers, D. Anguelov, H.-C. Pang, and D. Koller. Object pose detection in range scan data. In *CVPR (2)*, pages 2445–2452, 2006. 1

[19] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. Torr. Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. Ieee, 2008. 1

[20] L. Schwarz, D. Mateus, V. Castañeda, and N. Navab. Manifold learning for tof-based human body tracking and activity recognition. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2010. 2

[21] L. Schwarz, D. Mateus, and N. Navab. Recognizing multiple human activities and tracking full-body pose in unconstrained environments. *Pattern Recognition*, 45(1):11–23, 2012. 1

[22] J. Shotton, A. Fitzgibbon, M. Cook, T. S. M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, volume 2, page 3, 2011. 1, 2, 3

[23] L. Sigal, A. O. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007. 1

[24] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 103–110. IEEE, 2012. 2

[25] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *CVPR*, 2008. 1, 2

[26] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, pages 238–245, Washington, DC, USA, 2006. IEEE Computer Society. 1, 2

[27] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1359–1367, 2011. 1, 2